Consistent Pseudo-Maximum Likelihood Estimators and Groups of Transformations

Gouriéroux, Christian and Monfort, Alain and Zakoian, Jean-Michel

1University of Toronto, TSE and PSL, Dauphine, CREST, CREST

2018

Consistent Pseudo-Maximum Likelihood Estimators and Groups of Transformations

C., Gouriéroux[(1)], A. Monfort[(2)], and J.M., Zakoian[(3)]

(Revised version, June, 2018)

Abstract

In a transformation model $\boldsymbol{y}_t = c[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\beta}), \boldsymbol{u}_t]$, where the errors $\boldsymbol{u}_t$ are i.i.d and independent of the explanatory variables $\boldsymbol{x}_t$, the parameters can be estimated by a pseudo-maximum likelihood (PML) method, that is, by using a misspecified distribution of the errors, but the PML estimator of $\boldsymbol{\beta}$ is in general not consistent. We explain in this paper how to nest the initial model in an identified augmented model with more parameters in order to derive consistent PML estimators of appropriate functions of parameter $\boldsymbol{\beta}$. The usefulness of the consistency result is illustrated by examples of systems of nonlinear equations, conditionally heteroskedastic models, stochastic volatility, or models with spatial interactions.

Keywords : Pseudo-Maximum Likelihood, Transformation Model, Identification, Consistency, Stochastic Volatility, Conditional Heteroskedasticity, Spatial Interactions.

Forthcoming in *Econometrica*.

[1]University of Toronto, TSE and PSL, Dauphine.
[2]CREST.
[3]CREST, and University of Lille.
We thank the Editor and three anonymous referees for helpful comments and suggestions. We gratefully acknowledge financial support of the Chair ACPR: "Regulation and Systemic Risk", and of the chair LCL: "New Challenges for New Data".

1 Introduction

In this paper, we are interested in transformation models of the type:

$$\boldsymbol{y}_t = \boldsymbol{c}[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\beta}); \boldsymbol{u}_t], \tag{1.1}$$

where $\boldsymbol{y}_t \in \mathcal{Y}$ is a n-dimensional vector of observed endogenous variables, $\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\beta}) \in \mathcal{A} \subset \mathbb{R}^J$ is a vector of index functions depending on exogenous and/or lagged endogenous explanatory variables $\boldsymbol{x}_t$ and on a parameter $\boldsymbol{\beta} \in \mathcal{B} \subset \mathbb{R}^K$, and $\boldsymbol{u}_t \in \mathcal{U}$ are independent identically distributed (i.i.d.) error terms of dimension n , $\boldsymbol{u}_t$ being independent of $\boldsymbol{x}_t$. The true value of parameter $\boldsymbol{\beta} \in \mathcal{B}$ is $\boldsymbol{\beta}_0$, and the true distribution of the errors is $P_0 \in \mathcal{P}$. When P_0 is unknown, for given observations $(\boldsymbol{y}_t, \boldsymbol{x}_t), t = 1, \ldots, T$, the parameter of interest $\boldsymbol{\beta}$ can be estimated by pseudo-maximum likelihood (PML), that is, by a maximum likelihood approach of (1.1) using a given distribution $P \in \mathcal{P}$ for the errors. Since $P \neq P_0$, the PML estimator $\hat{\boldsymbol{\beta}}_T$ of $\boldsymbol{\beta}$ is in general not consistent for $\boldsymbol{\beta}_0$.

In this paper, we introduce an augmented version of the initial model (1.1), for which a function of $\boldsymbol{\beta}$ can be consistently estimated by PML. For this purpose, we assume that (i) the set of functions $\mathcal{C} = \{\boldsymbol{c}_{\boldsymbol{a}} : \boldsymbol{u} \mapsto \boldsymbol{c}[\boldsymbol{a}; \boldsymbol{u}], \boldsymbol{a} \in \mathcal{A}\}$ from $\mathcal{U}$ to $\mathcal{Y}$ is a group for the operation $\circ$ of function composition[4]; (ii) the function $\boldsymbol{a} \mapsto \boldsymbol{c}_{\boldsymbol{a}}$ is one-to-one from $\mathcal{A}$ to $\mathcal{C}$. Under these assumptions, the group structure on $\mathcal{C}$ induces a group structure on $\mathcal{A}$, for a group operation denoted $*$ which is defined by:

$$\boldsymbol{c}_{\boldsymbol{a}} \circ \boldsymbol{c}_{\boldsymbol{b}} = \boldsymbol{c}_{\boldsymbol{a}*\boldsymbol{b}}, \quad \boldsymbol{a}, \boldsymbol{b} \in \mathcal{A}.$$

The *augmented* transformation model is defined by:

$$\boldsymbol{y}_t = \boldsymbol{c}[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\beta}) * \boldsymbol{\lambda}; \boldsymbol{u}_t], \quad (\boldsymbol{\beta}, \boldsymbol{\lambda}) \in \mathcal{B} \times \mathcal{A}, \tag{1.2}$$

where the extra parameter $\boldsymbol{\lambda}$ varies freely in $\mathcal{A}$. This extra parameter plays the role of an "intercept" introduced at an appropriate place. Two cases can be considered:

1) when $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ is identifiable, we prove in this paper that the PML estimator of $(\boldsymbol{\theta}, \boldsymbol{\lambda}) := (\boldsymbol{\beta}, \boldsymbol{\lambda})$ is such that the PML estimator of $\boldsymbol{\theta} = \boldsymbol{\beta}$ converges to $\boldsymbol{\beta}_0$.

2) However, the initial model can already include some "intercept" parameters. In this case, we can put aside these intercepts $\boldsymbol{\lambda}(\boldsymbol{\beta})$, say, such that:

$$\boldsymbol{c}[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\beta}); \boldsymbol{u}_t] = \boldsymbol{c}[\overline{\boldsymbol{a}}(\boldsymbol{x}_t, \boldsymbol{\theta}(\boldsymbol{\beta})) * \lambda(\boldsymbol{\beta}); \boldsymbol{u}_t], \tag{1.3}$$

for some functions $\boldsymbol{\theta} : \mathcal{B} \to \boldsymbol{\theta}(\mathcal{B})$, $\boldsymbol{\lambda} : \mathcal{B} \to \mathcal{A}$ and $\overline{\boldsymbol{a}}$ valued in $\mathcal{A}$.
Then, the *identifiable augmented* model becomes:

$$\boldsymbol{y}_t = \boldsymbol{c}[\overline{\boldsymbol{a}}(\boldsymbol{x}_t, \boldsymbol{\theta}) * \boldsymbol{\lambda}; \boldsymbol{u}_t], \quad (\boldsymbol{\theta}, \boldsymbol{\lambda}) \in \boldsymbol{\theta}(\mathcal{B}) \times \mathcal{A}, \tag{1.4}$$

and we prove that the PML estimator of $\boldsymbol{\theta}$ is a consistent estimator of $\boldsymbol{\theta}(\boldsymbol{\beta}_0)$, while the PML estimator of $\boldsymbol{\lambda}$ does not necessarily converge to $\boldsymbol{\lambda}(\boldsymbol{\beta}_0)$. To get identification in model (1.4), the function $\boldsymbol{\beta} \mapsto \boldsymbol{\theta}(\boldsymbol{\beta})$ has often reduced rank: $\dim[\boldsymbol{\theta}(\mathcal{B})] \leq K = \dim(\mathcal{B})$, but with a possibly increased dimension of the augmented parameter: $\dim[\boldsymbol{\theta}(\mathcal{B})] + J \geq K = \dim(\mathcal{B})$. The parametrizations $\boldsymbol{\beta}$ and $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ are not necessarily one-to-one.

[4] A group $\mathcal{A}$ is a (finite or infinite) set of elements endowed with a binary operation $*$ (called the group operation) that satisfies the properties of closure (for each $a, b \in \mathcal{A}$, we have $a * b \in \mathcal{A}$), associativity ($(a * b) * c = a * (b * c)$), the identity property (existence of an *identity* element e such that $a * e = e * a = a$ for any $a \in \mathcal{A}$), and the inverse property (each $a \in \mathcal{A}$ admits an inverse $a^{-1} \in \mathcal{A}$ satisfying $a * a^{-1} = a^{-1} * a = e$).

Our analysis differs from the standard literature considering the consistency of the PML estimator of an index function interpretable as a conditional expectation, a conditional median and/or a conditional variance [see Gouriéroux et al. (1984), Bollerslev and Wooldridge (1992), Gouriéroux and Monfort (1995), Chapter 8]. It also differs from analyses where the shape of the true and/or pseudo error distribution is constrained. Due to the lack of ex-ante interpretation of the index in model (1.1), and to the lack of restrictions on the pseudo-distribution, we introduce a parameter $\boldsymbol{\lambda}$ allowing for consistency of PML estimators of $\boldsymbol{\theta}(\boldsymbol{\beta}_0)$ for any pseudo-distribution satisfying minimal regularity conditions. Our analysis is thus not limited to a class such as the linear exponential family [as in Gouriéroux et al. (1984)], the conditionally heteroskedastic models [as in Francq et al. (2011), Fan et al. (2014)], or to multivariate conditionally heteroskedastic dynamic regression models [as in Fiorentini, Sentana (2016)]. Section 2 illustrates the usefulness of the consistency result by considering the examples of regression model with conditional heteroskedasticity, Cholesky ARCH model, and model with homogenous spatial interactions. Section 3 derives the main result of the paper, that is, the consistency of the PML estimator. In the rest of the paper, we focus on linear (affine) transformation models. These models are studied in Section 4. We discuss the choice of the appropriate parametrization for the examples of Section 2 and provide other applications to network models, to the multivariate regression model with conditional heteroskedasticity, and to models for observations of volatility matrices. Assumptions and identification issues are discussed in Section 5, with special focus on the Cholesky ARCH model and on an interaction model. Section 6 derives the asymptotic distribution of the PML estimator for a class of linear commutative transformation models. Section 7 concludes. Additional results and proofs are collected in an on-line Appendix.

2 Examples of groups and augmented models

In this section, we consider examples of models (1.1) and (1.2) and their associated groups. At this stage, we discuss some identifiability issues in the augmented model (but the main discussion of identifiability is postponed to Section 5).

Example 1: The Newey-Steigerwald model

The model of interest is a unidimensional "regression" model with conditional heteroskedasticity:

$$y_t = a_1(\boldsymbol{x}_t, \boldsymbol{\beta}) + a_2(\boldsymbol{x}_t, \boldsymbol{\beta})u_t, \tag{2.1}$$

where $y_t, a_1(\cdot) \in \mathbb{R}$ and $a_2(\cdot) \in \mathbb{R}^* = \mathbb{R} - \{0\}$. The group of functions is $\mathcal{C} = \{c_a : u \mapsto a_1 + a_2 u, (a_1, a_2) \in \mathbb{R} \times \mathbb{R}^*\}$ and the operation on the associated group $\mathbb{R} \times \mathbb{R}^*$ is defined by $(a_1, a_2) * (b_1, b_2) = (a_1 + a_2 b_1, a_2 b_2)$. Note that this group is not commutative. Therefore, the augmented model is:

$$y_t = a_1(\boldsymbol{x}_t, \boldsymbol{\beta}) + \lambda_1 a_2(\boldsymbol{x}_t, \boldsymbol{\beta}) + \lambda_2 a_2(\boldsymbol{x}_t; \boldsymbol{\beta})u_t. \tag{2.2}$$

The introduction of two additional parameters λ_1, λ_2 to get consistency of PML estimators for Model (2.1) is due to Newey and Steigerwald (1997). The identifiability issue in Model (2.2) depends on functions a_1, a_2. For illustration purposes, let us consider the standard modelling with $a_1(\boldsymbol{x}_t, \boldsymbol{\beta}) = \boldsymbol{x}_t'\boldsymbol{\beta}_{11} + \boldsymbol{\beta}_{12}, a_1(\boldsymbol{x}_t, \boldsymbol{\beta}) = \exp(\boldsymbol{x}_t'\boldsymbol{\beta}_{21} + \boldsymbol{\beta}_{22})$, assuming that the regressors are not colinear. The augmented model is:

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_{11} + \boldsymbol{\beta}_{12} + \lambda_1 \exp(\boldsymbol{x}_t'\boldsymbol{\beta}_{21} + \boldsymbol{\beta}_{22}) + \lambda_2 \exp(\boldsymbol{x}_t'\boldsymbol{\beta}_{21} + \boldsymbol{\beta}_{22})u_t.$$

We see that the parameters λ_1, λ_2 and β_{22} are not identifiable and that the identified augmented model is

$$y_t \quad = \quad \boldsymbol{x}_t'\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2 + \lambda_1\lambda_2 \exp(\boldsymbol{x}_t'\boldsymbol{\theta}_3) + \lambda_2 \exp(\boldsymbol{x}_t'\boldsymbol{\theta}_3)u_t.$$

In fact, the initial model already includes an intercept on the log-volatility at the right place and corresponds to

$$\boldsymbol{\theta}(\boldsymbol{\beta}) = (\boldsymbol{\beta}_{11}', \boldsymbol{\beta}_{12}', \boldsymbol{\beta}_{21}')' := (\boldsymbol{\theta}_1'(\boldsymbol{\beta}), \boldsymbol{\theta}_2'(\boldsymbol{\beta}), \boldsymbol{\theta}_3'(\boldsymbol{\beta}))', \quad \boldsymbol{\lambda}(\boldsymbol{\beta}) = (0, \exp(\beta_{22}))'.$$

However, the usual intercept for the mean is not appropriate and, to recover consistency, a "risk premium" $\lambda_1\lambda_2 \exp(\boldsymbol{x}_t'\boldsymbol{\theta}_3)$ has to be introduced. Then the parameters $\boldsymbol{\beta}_{11}, \boldsymbol{\beta}_{12}$ and $\boldsymbol{\beta}_{21}$ can be consistently estimated by PML, but not parameter $\boldsymbol{\beta}_{22}$.

Example 2: Cholesky ARCH model

Let us consider a bivariate ($n = 2$) Cholesky ARCH model (see Dellaportas, Pourahmadi (2012)):

$$\boldsymbol{y}_t = \begin{pmatrix} a_{11}(\boldsymbol{x}_t, \boldsymbol{\beta}) & 0 \\ a_{21}(\boldsymbol{x}_t, \boldsymbol{\beta}) & a_{22}(\boldsymbol{x}_t, \boldsymbol{\beta}) \end{pmatrix} \boldsymbol{u}_t := \boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\beta})\boldsymbol{u}_t.$$

The group of functions is $\mathcal{C} = \{\boldsymbol{u} \mapsto \boldsymbol{A}\boldsymbol{u}, \boldsymbol{A} \text{ invertible lower triangular matrix}\}$. We can take[5] $\boldsymbol{a} = \text{vech}(\boldsymbol{A}) = (a_{11}, a_{21}, a_{22})'$ and the operation $*$ in the group of vectors in $\mathbb{R}^* \times \mathbb{R} \times \mathbb{R}^*$ is given by $\boldsymbol{a} * \boldsymbol{b} = \text{vech}(\boldsymbol{AB})$, that is $\boldsymbol{a} * \boldsymbol{b} = (a_{11}b_{11}, a_{21}b_{11} + a_{22}b_{21}, a_{22}b_{22})'$. The augmented model is:

$$\boldsymbol{y}_t = \begin{pmatrix} a_{11}(\boldsymbol{x}_t, \boldsymbol{\beta}) & 0 \\ a_{21}(\boldsymbol{x}_t, \boldsymbol{\beta}) & a_{22}(\boldsymbol{x}_t, \boldsymbol{\beta}) \end{pmatrix} \begin{pmatrix} \lambda_{11} & 0 \\ \lambda_{21} & \lambda_{22} \end{pmatrix} \boldsymbol{u}_t = \begin{pmatrix} a_{11}(\boldsymbol{x}_t, \boldsymbol{\beta})\lambda_{11} & 0 \\ a_{21}(\boldsymbol{x}_t, \boldsymbol{\beta})\lambda_{11} + a_{22}(\boldsymbol{x}_t, \boldsymbol{\beta})\lambda_{21} & a_{22}(\boldsymbol{x}_t, \boldsymbol{\beta})\lambda_{22} \end{pmatrix} \boldsymbol{u}_t.$$

Identifiability issues for this augmented model will be discussed in Section 5.

Example 3: Model with homogenous dynamic spatial interactions

The model of interest is:

$$\boldsymbol{y}_t = a_{1t} \begin{pmatrix} 1 & a_{2t} \dots & & a_{2t} \\ a_{2t} & 1 & & \vdots \\ \vdots & & \ddots & a_{2t} \\ a_{2t} & \dots & & 1 \end{pmatrix} \boldsymbol{u}_t, \tag{2.3}$$

with two index functions $a_{1t} = a_1(\boldsymbol{x}_t, \boldsymbol{\beta}) \in \mathbb{R}^*, a_{2t} = a_2(\boldsymbol{x}_t, \boldsymbol{\beta}) \in (-1/(n-1), 1)$, and $\boldsymbol{y}_t, \boldsymbol{u}_t \in \mathbb{R}^n$.

The group of functions for the composition operation is the set of functions:

$$\mathcal{C} = \left\{ \boldsymbol{c}_{\boldsymbol{a}}, \boldsymbol{a} = (a_1, a_2) \in \mathbb{R}^* \times \left(-\frac{1}{n-1}, 1 \right) \right\} \quad \text{where} \quad \boldsymbol{c}(\boldsymbol{a}, \boldsymbol{u}) = a_1 \begin{pmatrix} 1 & a_2 \dots & & a_2 \\ a_2 & 1 & & \vdots \\ \vdots & & \ddots & a_2 \\ a_2 & \dots & & 1 \end{pmatrix} \boldsymbol{u},$$

[5]The vech operator stacks into a vector the lower triangular part of a matrix. It is used in our framework for a non symmetric matrix.

and the operation $*$ of the associated group on $\mathcal{A}$ is defined by:

$$(a_1, a_2) * (b_1, b_2) = \left(a_1 b_1 [1 + (n-1) a_2 b_2], \frac{a_2 + b_2 + (n-2) a_2 b_2}{1 + (n-1) a_2 b_2} \right).$$

The identity element of this group is $\boldsymbol{e} = (1, 0)$ and the inverse of $\boldsymbol{a}$ is:

$$\boldsymbol{a}^{-1} = \left(\frac{1 + (n-2) a_2}{a_1 (1 - a_2)(1 + (n-1) a_2)}, \frac{-a_2}{1 + (n-2) a_2} \right),$$

which belongs to $\mathcal{A}$ for any $\boldsymbol{a} \in \mathcal{A}$. In this case, both groups $(\mathcal{A}, *)$ and $(\mathcal{C}, \circ)$ are Abelian.

The augmented model is:

$$\boldsymbol{y}_t = \boldsymbol{c}[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\beta}) * \boldsymbol{\lambda}, \boldsymbol{u}_t] = \tilde{a}_{1t} \begin{pmatrix} 1 & \tilde{a}_{2t} \dots & & \tilde{a}_{2t} \\ \tilde{a}_{2t} & 1 & & \vdots \\ \vdots & & \ddots & \tilde{a}_{2t} \\ \tilde{a}_{2t} & \dots & & 1 \end{pmatrix} \boldsymbol{u}_t,$$

with

$$\tilde{a}_{1t} = a_1(\boldsymbol{x}_t, \boldsymbol{\beta}) \lambda_1 [1 + (n-1) a_2(\boldsymbol{x}_t, \boldsymbol{\beta}) \lambda_2], \quad \tilde{a}_{2t} = \frac{a_2(\boldsymbol{x}_t, \boldsymbol{\beta}) + \lambda_2 + (n-2) a_2(\boldsymbol{x}_t, \boldsymbol{\beta}) \lambda_2}{1 + (n-1) a_2(\boldsymbol{x}_t, \boldsymbol{\beta}) \lambda_2}.$$

Identifiability of the augmented model will be discussed in Section 5.

3 Main result

We consider the model of interest

$$\boldsymbol{y}_t = \boldsymbol{c}[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\beta}); \boldsymbol{u}_t], \tag{3.1}$$

with true parameter value $\boldsymbol{\beta}_0$. The assumptions given in Section 1 are summarized and completed as follows.

Assumption A.1: $\boldsymbol{y}_t, \boldsymbol{u}_t \in \mathcal{Y} = \mathcal{U} \subset \mathbb{R}^n$, $\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\beta}) \in \mathcal{A} \subset \mathbb{R}^J$, $\boldsymbol{\beta} \in \mathcal{B} \subset \mathbb{R}^K$. Moreover,

(i) the set of functions $\mathcal{C} = \{\boldsymbol{c}_{\boldsymbol{a}} : \boldsymbol{u} \mapsto \boldsymbol{c}[\boldsymbol{a}; \boldsymbol{u}], \boldsymbol{a} \in \mathcal{A}\}$ is a group for the composition $\circ$;

(ii) the function $\boldsymbol{a} \mapsto \boldsymbol{c}_{\boldsymbol{a}}$ is one-to-one from $\mathcal{A}$ to $\mathcal{C}$.

(iii) $\mathcal{U}$ is a manifold, and, for any $\boldsymbol{a} \in \mathcal{A}$, the function $\boldsymbol{c}_{\boldsymbol{a}}$ is a diffeomorphism[6] from $\mathcal{U} \subset \mathbb{R}^n$ to $\mathcal{U}$.

As already mentioned, it follows from Assumption **A.1** that the group $(\mathcal{A}, *)$ induced by the group $(\mathcal{C}, \circ)$ is characterized by:

$$\boldsymbol{c}_{\boldsymbol{a}} \circ \boldsymbol{c}_{\boldsymbol{b}} = \boldsymbol{c}_{\boldsymbol{a} * \boldsymbol{b}}, \quad \boldsymbol{a}, \boldsymbol{b} \in \mathcal{A}, \tag{3.2}$$

and in the following, we denote by $\boldsymbol{e}$ the identity element and $\boldsymbol{a}^{-1}$ the inverse of $\boldsymbol{a}$ for the operation $*$: we have $\boldsymbol{c}[\boldsymbol{a} * \boldsymbol{a}^{-1}, \boldsymbol{u}] = \boldsymbol{c}[\boldsymbol{a}^{-1} * \boldsymbol{a}, \boldsymbol{u}] = \boldsymbol{c}[\boldsymbol{e}, \boldsymbol{u}] = \boldsymbol{u}$. We have $\boldsymbol{c}_{\boldsymbol{a}^{-1}} = \boldsymbol{c}_{\boldsymbol{a}}^{-1}$; therefore the sets $\mathcal{Y}$ and $\mathcal{U}$ must coincide. Note also that $\boldsymbol{y} = \boldsymbol{c}(\boldsymbol{a}, \boldsymbol{u}) \Leftrightarrow \boldsymbol{u} = \boldsymbol{c}(\boldsymbol{a}^{-1}, \boldsymbol{y})$.

[6] A diffeomorphism is a bijection which is differentiable and such that its inverse is differentiable as well.

By differentiating (3.2) with respect to $\boldsymbol{u}$, we get:

$$\frac{\partial \boldsymbol{c_a}}{\partial \boldsymbol{u}'} \circ \boldsymbol{c_b} = \frac{\partial \boldsymbol{c_{a*b}}}{\partial \boldsymbol{u}'} \cdot \left[\frac{\partial \boldsymbol{c_b}}{\partial \boldsymbol{u}'}\right]^{-1}, \quad \boldsymbol{a}, \boldsymbol{b} \in \mathcal{A}. \tag{3.3}$$

For expository purpose, we do not distinguish below functions a and $\overline{a}$ and we consider the identifiable (see assumptions below) augmented model:

$$\boldsymbol{y}_t = \boldsymbol{c}[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\theta}) * \boldsymbol{\lambda}; \boldsymbol{u}_t], \qquad (\boldsymbol{\theta}, \boldsymbol{\lambda}) \in \boldsymbol{\theta}(\mathcal{B}) \times \mathcal{A}, \tag{3.4}$$

Recall that the model of interest (3.1) and the augmented model (3.4) are linked via $\boldsymbol{c}[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\beta}); \boldsymbol{u}_t] = \boldsymbol{c}[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\theta}(\boldsymbol{\beta})) * \lambda(\boldsymbol{\beta}); \boldsymbol{u}_t]$. This model can be estimated by a PML method in which the errors are assumed to follow a given distribution P. This pseudo-distribution is in general different from the true distribution P_0. The PML estimator $(\hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\lambda}}_T)$ of the true value $(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0) = (\boldsymbol{\theta}(\boldsymbol{\beta}_0), \boldsymbol{\lambda}(\boldsymbol{\beta}_0))$ is defined as any measurable solution of:

$$(\hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\lambda}}_T) = \arg \max_{(\boldsymbol{\theta}, \boldsymbol{\lambda}) \in \boldsymbol{\theta}(\mathcal{B}) \times \mathcal{A}} L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}), \qquad \text{where} \quad L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \frac{1}{T} \sum_{t=1}^{T} l[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\theta}) * \boldsymbol{\lambda}, \boldsymbol{y}_t], \tag{3.5}$$

and l denotes the pseudo log-likelihood corresponding to the observation of date t.

Let us introduce the following assumptions:

Assumption A.2: The joint process $(\boldsymbol{x}_t, \boldsymbol{u}_t)$ is strictly stationary ergodic and $\boldsymbol{u}_t$ are i.i.d. error terms, with continuous true distribution P_0 and density g_0 w.r.t. the Lebesgue measure.

Assumption A.3: $\boldsymbol{x}_t$ and $\boldsymbol{u}_t$ are independent.

Assumption **A.3** concerns the variables $\boldsymbol{x}_t, \boldsymbol{u}_t$, not the processes $(\boldsymbol{x}_t), (\boldsymbol{u}_t)$. Let $E_{x,0}$ denote the expectation with respect to the stationary distribution of $(\boldsymbol{u}_t, \boldsymbol{x}_t)$

Assumption A.4: the function $l[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\theta}) * \boldsymbol{\lambda}, \boldsymbol{y}_t]|$ is measurable with respect to $(\boldsymbol{x}_t, \boldsymbol{y}_t)$ and $E_{x,0}|\sup_{(\boldsymbol{\theta}, \boldsymbol{\lambda}) \in \boldsymbol{\theta}(\mathcal{B}) \times \mathcal{A}} l[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\theta}) * \boldsymbol{\lambda}, \boldsymbol{y}_t]| < \infty$, Moreover, the random function $(\boldsymbol{\theta}, \boldsymbol{\lambda}) \to l[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\theta}) * \boldsymbol{\lambda}, \boldsymbol{y}_t]$ is a.s. continuous over $\boldsymbol{\theta}(\mathcal{B}) \times \mathcal{A}$.

We will consider a pseudo distribution with pseudo density g w.r.t. the Lebesgue measure on $\mathbb{R}^n$. Under Assumptions **A.1-A.4**, the objective function converges to a limiting objective function equal to an expected pseudo log-likelihood. More precisely, let E_0 denote the expectation with respect to the true errors distribution P_0, and let

$$L(\boldsymbol{a}; P, P_0) = E_0\{\log g[\boldsymbol{c}(\boldsymbol{a}^{-1}; \boldsymbol{u})] + \log|\det \frac{\partial \boldsymbol{c}}{\partial \boldsymbol{u}'}(\boldsymbol{a}^{-1}; \boldsymbol{u})|\}.$$

We now derive the expression of the limiting objective function $\lim_{T\to\infty} a.s. L_T(\boldsymbol{\theta}, \boldsymbol{\lambda})$ by applying the standard change of variables in multiple integrals based on the determinant of the Jacobian matrix.

Proposition 1 *Let us consider the nonlinear transformation model (3.4) with the true parameter value* $(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0)$. *Under Assumptions* **A.1-A.4**, *we have, for some constant* K *independent of* $\boldsymbol{\theta}, \boldsymbol{\lambda}$:

$$\lim_{T\to\infty} a.s. \frac{1}{T} \sum_{t=1}^{T} l[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\theta}) * \boldsymbol{\lambda}, \boldsymbol{y}_t] = E_x L[\boldsymbol{\lambda}_0^{-1} * \boldsymbol{a}^{-1}(\boldsymbol{x}_t, \boldsymbol{\theta}_0) * \boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\theta}) * \boldsymbol{\lambda}; P, P_0] + K.$$

Proof: We get:

$$\begin{aligned}
& \lim_{T\to\infty} a.s \frac{1}{T}\sum_{t=1}^{T} l[\boldsymbol{a}(\boldsymbol{x}_t,\boldsymbol{\theta}) * \boldsymbol{\lambda}, \boldsymbol{y}_t] \\
= \ & E_{x,0}\{\log g[\boldsymbol{c}(\boldsymbol{\lambda}^{-1} * \boldsymbol{a}^{-1}(\boldsymbol{x}_t,\boldsymbol{\theta}); \boldsymbol{y}_t)] + \log|\det[\frac{\partial \boldsymbol{c}}{\partial \boldsymbol{u}'}(\boldsymbol{\lambda}^{-1} * \boldsymbol{a}^{-1}(\boldsymbol{x}_t,\boldsymbol{\theta}); \boldsymbol{y}_t)]|\} \\
= \ & E_x E_0\{\log g[\boldsymbol{c}(\boldsymbol{\lambda}^{-1} * \boldsymbol{a}^{-1}(\boldsymbol{x}_t,\boldsymbol{\theta}); \boldsymbol{c}[\boldsymbol{a}(\boldsymbol{x}_t,\boldsymbol{\theta}_0) * \boldsymbol{\lambda}_0; \boldsymbol{u}_t])] \\
& + \log|\det[\frac{\partial \boldsymbol{c}}{\partial \boldsymbol{u}'}(\boldsymbol{\lambda}^{-1} * \boldsymbol{a}^{-1}(\boldsymbol{x}_t,\boldsymbol{\theta}); \boldsymbol{c}[\boldsymbol{a}(\boldsymbol{x}_t,\boldsymbol{\theta}_0) * \boldsymbol{\lambda}_0; \boldsymbol{u}_t])]|\} \\
= \ & E_x E_0\{\log g[\boldsymbol{c}(\boldsymbol{\lambda}^{-1} * \boldsymbol{a}^{-1}(\boldsymbol{x}_t,\boldsymbol{\theta}) * \boldsymbol{a}(\boldsymbol{x}_t,\boldsymbol{\theta}_0) * \boldsymbol{\lambda}_0; \boldsymbol{u}_t)] \\
& + \log|\det[\frac{\partial \boldsymbol{c}}{\partial \boldsymbol{u}'}(\boldsymbol{\lambda}^{-1} * \boldsymbol{a}^{-1}(\boldsymbol{x}_t,\boldsymbol{\theta}) * \boldsymbol{a}(\boldsymbol{x}_t,\boldsymbol{\theta}_0) * \boldsymbol{\lambda}_0; \boldsymbol{u}_t)]|\} + K \\
= \ & E_x L[\boldsymbol{\lambda}_0^{-1} * \boldsymbol{a}^{-1}(\boldsymbol{x}_t,\boldsymbol{\theta}_0) * \boldsymbol{a}(\boldsymbol{x}_t,\boldsymbol{\theta}) * \boldsymbol{\lambda}; P, P_0] + K, \qquad (3.6)
\end{aligned}$$

where $K = -E_x E_0 \log|\det[\frac{\partial \boldsymbol{c}}{\partial \boldsymbol{u}'}(\boldsymbol{a}(\boldsymbol{x}_t,\boldsymbol{\theta}_0) * \boldsymbol{\lambda}_0; \boldsymbol{u}_t)]|$ is independent of $\boldsymbol{\theta}, \boldsymbol{\lambda}$, and E_x denotes the expectation with respect to the stationary distribution of $\boldsymbol{x}_t$. The first equality follows from the ergodic theorem, which can be applied under **A.2** and **A.4**. The second equality uses the independence assumption **A.3**. The third equality uses (3.2) and the last one is obtained by applying (3.3).

QED

Then we make the following additional assumptions:

Assumption A.5: There is a unique solution to the optimization problem:

$$\boldsymbol{a}_0^* = \arg\max_{\boldsymbol{a}\in\mathcal{A}} L(\boldsymbol{a}; P, P_0).$$

Assumption A.6: There is a unique solution to the optimization problem:

$$(\boldsymbol{\theta}_0^*, \boldsymbol{\lambda}_0^*) = \arg\max_{(\boldsymbol{\theta},\boldsymbol{\lambda})\in\boldsymbol{\theta}(\mathcal{B})\times\mathcal{A}} E_x L[\boldsymbol{\lambda}_0^{-1} * \boldsymbol{a}^{-1}(\boldsymbol{x}_t,\boldsymbol{\theta}_0) * \boldsymbol{a}(\boldsymbol{x}_t,\boldsymbol{\theta}) * \boldsymbol{\lambda}; P, P_0]. \qquad (3.7)$$

Assumption A.7: $\boldsymbol{\theta}(\mathcal{B})$ and $\mathcal{A}$ are compact parameter sets.

Under Assumptions **A.1-A.7**, we will show that the PML estimator $(\hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\lambda}}_T)$ exists and tends a.s. to the pseudo-true value $(\boldsymbol{\theta}_0^*, \boldsymbol{\lambda}_0^*)$.

Proposition 2 *Let us consider the transformation model (3.4) with true parameter value* $(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0)$*. Under Assumptions* **A.1-A.7***, the PML estimator* $(\hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\lambda}}_T)$ *converges a.s. to* $\boldsymbol{\theta}_0^* = \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^* = \boldsymbol{\lambda}_0 * \boldsymbol{a}_0^*$.

Proof: Let us first prove that the limit criterion is uniquely maximized at $(\boldsymbol{\theta}_0^*, \boldsymbol{\lambda}_0^*)$. By Assumption **A.5** and the fact that $(\mathcal{A}, *)$ is a group, we get:

$$L[\boldsymbol{\lambda}_0^{-1} * \boldsymbol{a}^{-1}(\boldsymbol{x}_t,\boldsymbol{\theta}_0) * \boldsymbol{a}(\boldsymbol{x}_t,\boldsymbol{\theta}) * \boldsymbol{\lambda}; P, P_0] \leq L(\boldsymbol{a}_0^*; P, P_0), \quad \forall \boldsymbol{x}_t \in \mathcal{X}, \boldsymbol{\theta} \in \boldsymbol{\theta}(\mathcal{B}), \boldsymbol{\lambda} \in \mathcal{A}.$$

We deduce that:

$$E_x L[\boldsymbol{\lambda}_0^{-1} * \boldsymbol{a}^{-1}(\boldsymbol{x}_t,\boldsymbol{\theta}_0) * \boldsymbol{a}(\boldsymbol{x}_t,\boldsymbol{\theta}) * \boldsymbol{\lambda}; P, P_0] \leq L(\boldsymbol{a}_0^*; P, P_0),$$

and the maximum is reached for $\boldsymbol{\theta}_0$ and $\boldsymbol{\lambda}_0^*$ satisfying $\boldsymbol{\lambda}_0^{-1} * \boldsymbol{\lambda}_0^* = \boldsymbol{a}_0^*$. The uniqueness of the maximizer follows directly from **A.6.**

Now we will prove that, for any $(\boldsymbol{\theta}_1, \boldsymbol{\lambda}_1) \neq (\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*)$, there exists a neighborhood $V(\boldsymbol{\theta}_1, \boldsymbol{\lambda}_1)$ such that

$$\lim_{T\to\infty} a.s. \sup_{(\boldsymbol{\theta},\boldsymbol{\lambda})\in V(\boldsymbol{\theta}_1,\boldsymbol{\lambda}_1)} L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) < L(\boldsymbol{a}_0^*; P, P_0) + K, \tag{3.8}$$

where K is defined in (3.6). For any positive integer k, let $V_k(\boldsymbol{\theta}_1, \boldsymbol{\lambda}_1)$ be the open ball of center $(\boldsymbol{\theta}_1, \boldsymbol{\lambda}_1)$ and radius $1/k$. We have

$$\begin{aligned}\lim_{T\to\infty} a.s. \sup_{(\boldsymbol{\theta},\boldsymbol{\lambda})\in V_k(\boldsymbol{\theta}_1,\boldsymbol{\lambda}_1)} L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) &\leq \lim_{T\to\infty} a.s. \frac{1}{T}\sum_{t=1}^{T} \sup_{(\boldsymbol{\theta},\boldsymbol{\lambda})\in V_k(\boldsymbol{\theta}_1,\boldsymbol{\lambda}_1)} l[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\theta}) * \boldsymbol{\lambda}, \boldsymbol{y}_t] \\ &= E_x E_0 \sup_{(\boldsymbol{\theta},\boldsymbol{\lambda})\in V_k(\boldsymbol{\theta}_1,\boldsymbol{\lambda}_1)} l[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\theta}) * \boldsymbol{\lambda}, \boldsymbol{y}_t],\end{aligned}$$

by the ergodic theorem. By Beppo Levi's theorem, $E_x E_0 \sup_{(\boldsymbol{\theta},\boldsymbol{\lambda})\in V_k(\boldsymbol{\theta}_1,\boldsymbol{\lambda}_1)} l[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\theta}) * \boldsymbol{\lambda}, \boldsymbol{y}_t]$ decreases to $E_x E_0 l[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\theta}_1) * \boldsymbol{\lambda}_1, \boldsymbol{y}_t]$ as $k \to \infty$. We have already shown that the latter expectation is strictly less than $L(\boldsymbol{a}_0^*; P, P_0) + K$, thus (3.8) is established.

The end of the proof uses a standard compactness argument. First note that for any neighborhood $V(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*)$ of $(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*)$,

$$\lim_{T\to\infty} a.s. \sup_{V(\boldsymbol{\theta}_0,\boldsymbol{\lambda}_0^*)\cap\boldsymbol{\theta}(\mathcal{B})\times\mathcal{A}} L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) \geq \lim_{T\to\infty} a.s. \; L_T(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*) = L(\boldsymbol{a}_0^*; P, P_0) + K. \tag{3.9}$$

Next, we note that the compact set $\boldsymbol{\theta}(\mathcal{B}) \times \mathcal{A}$ is covered by the union of an arbitrary neighborhood $V(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*)$ of $(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*)$ and the set of the neighborhoods $V(\boldsymbol{\theta}_1, \boldsymbol{\lambda}_1)$ satisfying (3.8), with $(\boldsymbol{\theta}_1, \boldsymbol{\lambda}_1) \in \boldsymbol{\theta}(\mathcal{B}) \times \mathcal{A} \setminus V(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*)$. Thus, there exists a finite subcover of Θ of the form $V(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*), V(\boldsymbol{\theta}_1, \boldsymbol{\lambda}_1), \ldots, V(\boldsymbol{\theta}_k, \boldsymbol{\lambda}_k)$, where for $i = 1, \ldots, k$, $V(\boldsymbol{\theta}_i, \boldsymbol{\lambda}_i)$ satisfies (3.8). It follows that

$$\sup_{\boldsymbol{\theta}(\mathcal{B})\times\mathcal{A}} L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \max_{i=1,\ldots,k} \sup_{V(\boldsymbol{\theta}_i,\boldsymbol{\lambda}_i)\cap\boldsymbol{\theta}(\mathcal{B})\times\mathcal{A}} L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}).$$

The inequalities (3.8) and (3.9) show that, almost surely, the PML estimator belongs to $V(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*)$ for T large enough. The consistency of the PML estimator follows.

QED

At this stage, it is useful to introduce the notion of *generic model* defined as $\tilde{\boldsymbol{y}} = \boldsymbol{c}(\boldsymbol{a}; \boldsymbol{u})$, with the true parameter $\boldsymbol{a}_0 = \boldsymbol{e}$, and the true error distribution P_0. The objective function $L(\boldsymbol{a}; P, P_0)$ can be interpreted as a limiting pseudo log-likelihood in this model without explanatory variable. Indeed, we have, for any $\boldsymbol{a} \in \mathcal{A}$, $\tilde{\boldsymbol{y}} \in \mathcal{U} \subset \mathbb{R}^n$,

$$\begin{aligned}E_0 l(\boldsymbol{a}, \tilde{\boldsymbol{y}}) &= E_0\{\log g[\boldsymbol{c}(\boldsymbol{a}^{-1}; \tilde{\boldsymbol{y}})] + \log|\det \frac{\partial \boldsymbol{c}}{\partial \boldsymbol{u}'}(\boldsymbol{a}^{-1}; \tilde{\boldsymbol{y}})|\} \\ &= E_0\{\log g[\boldsymbol{c}(\boldsymbol{a}^{-1}; \boldsymbol{u})] + \log|\det \frac{\partial \boldsymbol{c}}{\partial \boldsymbol{u}'}(\boldsymbol{a}^{-1}; \boldsymbol{u})|\} \\ &= L(\boldsymbol{a}; P, P_0),\end{aligned}$$

since, for the true value of the parameter, $\tilde{\boldsymbol{y}} = \boldsymbol{c}(\boldsymbol{a}_0; \boldsymbol{u}) = \boldsymbol{c}(\boldsymbol{e}; \boldsymbol{u}) = \boldsymbol{u}$.

Remark 1: It is usual in practice to introduce the effect of the explanatory variables through some parameters. In other words, a generic model without explanatory variables, $\tilde{\boldsymbol{y}} = \boldsymbol{c}(\boldsymbol{a}; \boldsymbol{u})$, is transformed into an econometric model as $\boldsymbol{y}_t = \boldsymbol{c}[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\beta}); \boldsymbol{u}_t]$, say. In general, a PML estimator of $\boldsymbol{\beta}_0$ in the econometric model is not consistent. Proposition 2 means that an artificial extra parameter $\boldsymbol{\lambda}$ and possibly a (re)parametrization have to be introduced in the model to ensure the consistency of the PML estimator of some functions of $\boldsymbol{\beta}_0$ (see the Introduction and Section 5).

Remark 2: In a semi non-parametric estimation approach of Model (3.4), the vector of unknown parameters becomes $(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0, P_0) = (\boldsymbol{\theta}_0(\boldsymbol{\beta}_0), \boldsymbol{\lambda}_0(\boldsymbol{\beta}_0), P_0)$. However, it can be noted that $\boldsymbol{\lambda}_0$ and P_0 may not be identifiable, whatever the estimation method. Indeed, at the true parameter value, equation (3.4) can equivalently be written as:

$$\boldsymbol{y}_t = \boldsymbol{c}[\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\theta}_0); \boldsymbol{v}_t], \qquad \boldsymbol{v}_t = \boldsymbol{c}(\boldsymbol{\lambda}_0; \boldsymbol{u}_t). \tag{3.10}$$

The distribution of $\boldsymbol{y}_t$ conditional on $\boldsymbol{x}_t$ is thus invariant by any change of $(\boldsymbol{\lambda}_0, P_0)$ leaving unchanged the distribution of $\boldsymbol{c}(\boldsymbol{\lambda}_0; \boldsymbol{u}_t)$. However, the distribution of $\boldsymbol{v}_t$ is identifiable and can be estimated through residuals $\hat{\boldsymbol{v}}_{t,T} = \boldsymbol{c}[\boldsymbol{a}(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_T)^{-1}; \boldsymbol{y}_t]$, $t = 1, \ldots, T$.

Remark 3: Proposition 2 can be applied after preliminary one-to-one changes on the observed variables $\boldsymbol{y} \mapsto \boldsymbol{y}^* = \boldsymbol{h}_1(\boldsymbol{y}) \in \mathcal{Y}^* = \mathcal{U}^*$, errors $\boldsymbol{u} \mapsto \boldsymbol{u}^* = \boldsymbol{h}_2(\boldsymbol{u}) \in \mathcal{U}^* = \mathcal{Y}^*$, and/or indexes $\boldsymbol{a} \mapsto \boldsymbol{a}^* = \boldsymbol{h}_3(\boldsymbol{a})$, that is to econometric models of the form:

$$\boldsymbol{y}_t^* = \boldsymbol{h}_1 \left\{ \boldsymbol{c}[\boldsymbol{h}_3^{-1}(\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\beta})); \boldsymbol{h}_2^{-1}(\boldsymbol{u}_t^*)] \right\} \equiv \boldsymbol{c}^*[\boldsymbol{a}^*(\boldsymbol{x}_t, \boldsymbol{\beta}); \boldsymbol{u}_t^*].$$

The property of group of $(\mathcal{A}, *)$ is easily transferred to the new parametrization $\boldsymbol{a}^* \in \mathcal{A}^* = \boldsymbol{h}_3(\mathcal{A})$, with the operation $\tilde{*}$ defined by $\boldsymbol{a}^* \tilde{*}\, \boldsymbol{b}^* = \boldsymbol{h}_3[\boldsymbol{h}_3^{-1}(\boldsymbol{a}^*) * \boldsymbol{h}_3^{-1}(\boldsymbol{b}^*)]$.

4 Application to linear transformation models

In this section, we focus on linear (affine) transformation models and provide different examples, some of them including as special cases the examples of Section 2. As above, we do not distinguish below the functions a and $\overline{a}$ of the augmented and identifiable augmented models defined in (1.3).

4.1 Linear transformation model

Proposition 2 can in particular be applied to linear transformation models:

$$\boldsymbol{y}_t = \boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\beta})\boldsymbol{u}_t, \quad \boldsymbol{\beta} \in \mathcal{B}, \boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\beta}) \in \mathcal{A}, \tag{4.1}$$

where $\boldsymbol{y}_t, \boldsymbol{u}_t$ are $n \times 1$ vectors, $\boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\beta})$ is a $n \times n$ matrix and $\mathcal{A}$ is a group for the usual matrix multiplication operation, that is, a sub-group of the group of invertible matrices (by Cayley's Theorem). The associated identifiable augmented model is:

$$\boldsymbol{y}_t = \boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\theta})\boldsymbol{\Lambda}\boldsymbol{u}_t, \quad \boldsymbol{\theta} \in \boldsymbol{\theta}(\mathcal{B}), \boldsymbol{\Lambda} \in \mathcal{A}. \tag{4.2}$$

In this case, the generic model is $\tilde{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{u}$, with the true parameter $\boldsymbol{A}_0 = \boldsymbol{I}_n$, and the true error distribution P_0. We have, for any $\boldsymbol{A} \in \mathcal{A}$, $\tilde{\boldsymbol{y}} \in \mathbb{R}^n$,

$$E_0 l(\boldsymbol{A}, \tilde{\boldsymbol{y}}) = E_0 \log g(\boldsymbol{A}^{-1}\tilde{\boldsymbol{y}}) + \log|\det[\boldsymbol{A}^{-1}]| \;\; = \;\; E_0 \log g[\boldsymbol{A}^{-1}\boldsymbol{u}] + \log|\det[\boldsymbol{A}^{-1}]| = L(\boldsymbol{A}; P, P_0).$$

If $\mathcal{A}$ is an Abelian group, that is a commutative group, the augmented model (4.2) can equivalently be written as $\boldsymbol{y}_t = \boldsymbol{\Lambda A}(\boldsymbol{x}_t, \boldsymbol{\theta})\boldsymbol{u}_t$. When $\mathcal{A}$ is not Abelian, the augmented model has to be taken of the form (4.2).

Corollary 1 *Let us consider the linear transformation model (4.1). Under the assumptions of Proposition 2, we have:* $\boldsymbol{\theta}_0^* = \boldsymbol{\theta}_0, \boldsymbol{\Lambda}_0^* = \boldsymbol{\Lambda}_0 \boldsymbol{A}_0^*$*, where* $\boldsymbol{A}_0^* = \arg\max_{\boldsymbol{A}\in\mathcal{A}} L(\boldsymbol{A}; P, P_0)$.

4.2 Linear affine transformation model

The linear affine transformation model is:

$$\boldsymbol{y}_t = \boldsymbol{\mu}(\boldsymbol{x}_t, \boldsymbol{\beta}) + \boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\beta})\boldsymbol{u}_t, \quad \boldsymbol{\beta} \in \mathcal{B}, \boldsymbol{\mu} \in \mathcal{E}, \boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\beta}) \in \mathcal{A}. \tag{4.3}$$

We have: $\boldsymbol{c}(\boldsymbol{a}, \boldsymbol{u}) = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{u}$ where $\boldsymbol{a} = (\boldsymbol{\mu}, \boldsymbol{A})$, and the group operation is: $(\boldsymbol{\mu}_1, \boldsymbol{A}_1) * (\boldsymbol{\mu}_2, \boldsymbol{A}_2) = (\boldsymbol{\mu}_1 + \boldsymbol{A}_1\boldsymbol{\mu}_2, \boldsymbol{A}_1\boldsymbol{A}_2)$. The associated identified augmented model is:

$$\boldsymbol{y}_t = \boldsymbol{\mu}(\boldsymbol{x}_t, \boldsymbol{\theta}) + \boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\theta})\boldsymbol{\lambda} + \boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\theta})\boldsymbol{\Lambda}\boldsymbol{u}_t, \quad \boldsymbol{\theta} \in \boldsymbol{\theta}(\mathcal{B}), \boldsymbol{\lambda} \in \mathcal{E}, \boldsymbol{\Lambda} \in \mathcal{A}. \tag{4.4}$$

Then, Proposition 2 can be applied whenever $\mathcal{A}$ is a group for the multiplication of matrices, $\mathcal{E}$ is closed for the addition and such that $\mathcal{A}\mathcal{E} \subset \mathcal{E}$. This augmented model is a multivariate extension of the Newey, Steigerwald univariate model in Example 1.

We can associate with model (4.3) the generic model $\tilde{\boldsymbol{y}} = \boldsymbol{b} + \boldsymbol{A}\boldsymbol{u}$, with true parameters $\boldsymbol{b}_0 = \boldsymbol{0}, \boldsymbol{A}_0 = \boldsymbol{I}_n$, and true error distribution P_0. The limiting pseudo log-likelihood corresponding to the generic model is:

$$\begin{aligned} L(\boldsymbol{A}, \boldsymbol{b}; P, P_0) &= E_0 \log g[\boldsymbol{A}^{-1}(\tilde{\boldsymbol{y}} - \boldsymbol{b})] + \log|\det[\boldsymbol{A}^{-1}]| \\ &= E_0 \log g(-\boldsymbol{A}^{-1}\boldsymbol{b} + \boldsymbol{A}^{-1}\boldsymbol{u}) + \log|\det[\boldsymbol{A}^{-1}]|. \end{aligned}$$

Corollary 2 *Let us consider the linear affine transformation model (4.3). Under the assumptions of Proposition 2, we have:* $\boldsymbol{\theta}_0^* = \boldsymbol{\theta}_0, \boldsymbol{\Lambda}_0^* = \boldsymbol{\Lambda}_0\boldsymbol{A}_0^*$ *and* $\boldsymbol{\lambda}_0^* = \boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}_0\boldsymbol{b}_0^*$*, where* $(\boldsymbol{A}_0^*, \boldsymbol{b}_0^*) = \arg\max_{\boldsymbol{A}\in\mathcal{A}, \boldsymbol{b}\in\mathcal{E}} L(\boldsymbol{A}, \boldsymbol{b}; P, P_0)$.

When $\mathcal{E} = \{0\}$, we are back to Model (4.1). However, there exist structural models where the set $\mathcal{E}$ is constrained without being reduced to $\{0\}$. This is the case of the one-factor model used in the Arbitrage Pricing Theory (APT). The standard APT model is such that $\tilde{\boldsymbol{y}} = \mu\boldsymbol{f} + (b\boldsymbol{I}_n + c\boldsymbol{f}\boldsymbol{f}')\boldsymbol{u}$ with parameters μ, b, c, (the factorial direction $\boldsymbol{f}$ being fixed), where $\tilde{\boldsymbol{y}}$ is a vector of excess returns. In this setting, $\mathcal{E}$ is the vector space generated by $\boldsymbol{f}$.

Model (4.3) can be considered as a limit case of Model (4.1). Indeed, let us consider the set $\mathcal{A}^*$ of matrices $\begin{pmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{0} & 1 \end{pmatrix}$, where $\boldsymbol{A} \in \mathcal{A}$, $\boldsymbol{b} \in \mathcal{E}$ with $\boldsymbol{A}\mathcal{E} \subset \mathcal{E}$. Then $\mathcal{A}^*$ is also a group for the multiplication of matrices. A generic model is $\begin{pmatrix} \tilde{\boldsymbol{y}} \\ 1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{0} & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{u} \\ 1 \end{pmatrix}$, which is equivalent to $\tilde{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{u} + \boldsymbol{b}$. This is a limit case of the linear model in which the errors distribution is degenerate.

4.3 Exponential transformation models

We now provide specific examples of linear transformation models through their associated generic model. In the exponential transformation model, the matrix $\boldsymbol{A}$ is of the type:

$$\boldsymbol{A} = \exp\left(-\sum_{j=1}^{J} a_j \boldsymbol{C}_j\right), \tag{4.5}$$

where the $\boldsymbol{C}_j, j = 1, \ldots, J$ are $n \times n$ matrices that commute.[7] The augmented model is obtained with:

$$\boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\theta})\boldsymbol{\Lambda} = \exp\left(-\sum_{j=1}^{J}[a_j(\boldsymbol{x}_t, \boldsymbol{\theta}) + \lambda_j]\boldsymbol{C}_j\right).$$

The extra parameter can be interpreted as an intercept and the augmented model as the econometric model with intercept.

Example 3 (cont.): Model with homogenous spatial interactions

Let us consider a decomposition of $I\!R^n$ into J orthogonal vector spaces and let us denote $\boldsymbol{P}_j, j = 1, \ldots, J$, the associated orthogonal projectors. These projectors commute since $\boldsymbol{P}_j\boldsymbol{P}_k = 0, \forall j \neq k$, and it can be checked that:

$$\exp(-\sum_{j=1}^{J} a_j \boldsymbol{P}_j) = \sum_{j=1}^{J} \exp(-a_j)\boldsymbol{P}_j. \tag{4.6}$$

Then we can use Proposition 2 for the Abelian group of matrices $\exp(-\sum_{j=1}^{J} a_j \boldsymbol{C}_j)$ with $\boldsymbol{C}_j = \boldsymbol{P}_j$. Example 3 corresponds to the special case where: $J = 2$, $\boldsymbol{P}_1 = \mathbb{1}\mathbb{1}'/n, \boldsymbol{P}_2 = \boldsymbol{I}_n - \mathbb{1}\mathbb{1}'/n$, $\mathbb{1}$ denotes the vector with unitary components, and the transformation is of the type

$$\exp\{-\alpha_1(\boldsymbol{x}_t, \boldsymbol{\beta})\}\boldsymbol{P}_1 + \exp\{-\alpha_2(\boldsymbol{x}_t, \boldsymbol{\beta})\}\boldsymbol{P}_2,$$

where $\alpha_1(\boldsymbol{x}_t, \boldsymbol{\beta}), \alpha_2(\boldsymbol{x}_t, \boldsymbol{\beta})$ are in a one-to-one relationship with $a_1(\boldsymbol{x}_t, \boldsymbol{\beta}), a_2(\boldsymbol{x}_t, \boldsymbol{\beta})$ of Example 3 (see Section 5.3 below).

Example 4: Observations of volatility matrices

The set of symmetric positive definite (SPD) matrices is also a manifold denoted $\mathrm{Sym}^+(n)$ and our approach can be applied to SPD matrix valued data such as observed volatility matrices [see Arsigny et al. (2007), Yuan et al. (2012), Huang et al. (2014), and the references therein for other SPD matrix valued data in the medical imaging literature].

Let us consider a SPD matrix $\boldsymbol{U}$ of size (n, n), another (n, n) matrix $\boldsymbol{B}$, and the application:

$$\boldsymbol{U} \mapsto \exp(a\boldsymbol{B})\boldsymbol{U}\exp(a\boldsymbol{B}') := \boldsymbol{Y}.$$

[7] The exponential of a matrix is defined as:

$$\exp(a\boldsymbol{C}) = \sum_{h=0}^{\infty} \frac{(a^h \boldsymbol{C}^h)}{h!}.$$

This defines a group of linear transformations on $\text{Sym}^+(n)$. To link this group with our general specification, let us apply the vec operator. We have:

$$\text{vec}(\boldsymbol{Y}) = \text{vec}\{\exp(a\boldsymbol{B})\boldsymbol{U}\exp(a\boldsymbol{B}')\} = \{\exp(a\boldsymbol{B}) \otimes \exp(a\boldsymbol{B})\}\text{vec}(\boldsymbol{U}), \tag{4.7}$$

where $\otimes$ denotes the Kronecker product. The transformation in (4.7) can also be written as: $\text{vec}(\boldsymbol{Y}) = \exp(a\boldsymbol{C})\text{vec}(\boldsymbol{U})$ where $\boldsymbol{C}$ is the so-called infinitesimal generator.

To find matrix $\boldsymbol{C}$, let us now consider the behaviour of this transformation when a is close to zero. We get:

$$\{\exp(a\boldsymbol{B})\otimes\exp(a\boldsymbol{B})\}\text{vec}(\boldsymbol{U}) \sim \{(\boldsymbol{I}+a\boldsymbol{B})\otimes(\boldsymbol{I}+a\boldsymbol{B})\}\text{vec}(\boldsymbol{U}) \sim \text{vec}(\boldsymbol{U})+a\{\boldsymbol{B}\otimes\boldsymbol{I}+\boldsymbol{I}\otimes\boldsymbol{B}\}\text{vec}(\boldsymbol{U}).$$

Therefore, the underlying infinitesimal generator $\boldsymbol{C}$ is:

$$\boldsymbol{C} = \boldsymbol{B} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{B}, \tag{4.8}$$

and the transformation (4.7) can be equivalently written as:

$$\text{vec}(\boldsymbol{Y}) = \exp\{a(\boldsymbol{B} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{B})\}\text{vec}(\boldsymbol{U}). \tag{4.9}$$

Therefore, the methodology of our paper is valid for groups of such transformations applied to observed realized volatility matrices, or to observed implied volatility matrices derived from observed derivative prices.

Example 5: Stationary spatial models

Let us denote $\boldsymbol{C}(k)$ the (n, n) matrix with unitary values on the upper k^{th} diagonal and with zero anywhere else, i.e. with entries $C_{ij}(k) = 1$, if $j = i + k$, $= 0$, otherwise. By convention, $\boldsymbol{C}(k) = 0$ for $k \geq n$. These matrices commute since:

$$\boldsymbol{C}(k)\boldsymbol{C}(l) = \boldsymbol{C}(k + l) = \boldsymbol{C}(1)^{k+l}, \forall k, l \geq 0. \tag{4.10}$$

By applying formula (4.10), we see that:

$$\exp[\sum_{j=0}^{n-1} a_j \boldsymbol{C}(j)] \equiv \sum_{j=0}^{n-1} b_j(a)\boldsymbol{C}(j), \text{ say,} \tag{4.11}$$

where $b(a) = [b_j(a)]$ is a one-to-one function of the vector $a = (a_j)$. Formula (4.11) explains how to reparametrize for asymptotic bias adjustment spatial models of the type:

$$\tilde{\boldsymbol{y}} = \begin{pmatrix} b_0 & b_1 & b_{n-1} \\ 0 & \ddots & b_1 \\ 0 & 0 & b_0 \end{pmatrix} \boldsymbol{u} \equiv \boldsymbol{B}\boldsymbol{u}. \tag{4.12}$$

Such specifications are the spatial reduced form counterparts of models of the type $\tilde{\boldsymbol{y}} = \boldsymbol{R}\tilde{\boldsymbol{y}} + \boldsymbol{u}_t$, where $(\boldsymbol{I}_n - \boldsymbol{R})^{-1} = \boldsymbol{B}$, since the inverse of a triangular Toeplitz matrix is also triangular Toeplitz for large n. Thus, to facilitate bias adjustment, it is preferable to parametrize directly the reduced form, that is to avoid the modelling with simultaneous equations. These types of models are used in the literature on spatial data, networks, interactions, and peer effects. [8] Model (4.12) corresponds to hierarchical peer effects.

[8] See e.g. Cressie (1993), Banerjee et al. (2004) for spatial data, Manski (1993), Goldsmith-Pinkham and Imbens (2013) for linear-in-mean models with peer effects, Felsenstein (1973) for evolutionary trees in Genomics.

5 Identification

Let us discuss the assumptions **A.5-A.6** and the identification issues in the linear transformation model (4.2). First note that the assumptions depend on both P and P_0. For instance, Assumption **A.6** defined functions $\boldsymbol{\theta}_0^*(\boldsymbol{\theta}_0, \boldsymbol{\Lambda}_0; P, P_0), \boldsymbol{\Lambda}_0^*(\boldsymbol{\theta}_0, \boldsymbol{\Lambda}_0; P, P_0)$. For given P, P_0, we get the binding functions that explain how the pseudo-true values $\boldsymbol{\theta}_0^*, \boldsymbol{\Lambda}_0^*$ depend on the true values $\boldsymbol{\theta}_0, \boldsymbol{\Lambda}_0$. These binding functions extend in a semi-parametric framework the parametric binding function introduced in (White (1982), Gouriéroux et al. (1984)). We expect Assumptions **A.5-A.6** to be satisfied for a large class of true and pseudo distributions, i.e. for any $P, P_0 \in \mathcal{P}$. In particular they have to be satisfied for $P = P_0 \in \mathcal{P}$. When $P = P_0$, the PML reduces to the standard ML estimator, is consistent i.e. $\boldsymbol{\theta}_0^* = \boldsymbol{\theta}_0, \boldsymbol{\Lambda}_0^* = \boldsymbol{\Lambda}_0$, and Assumption **A.6** is simply the standard asymptotic identification condition.

Let us now discuss in more details Assumption **A.5** on the generic artificial model and Assumption **A.6** on the augmented model (4.2).

5.1 Identification in the generic model

In the generic model, $\tilde{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{u}$, with $\boldsymbol{A}_0 = \boldsymbol{I}_n$ and P_0 the true distribution, the pseudo-true value $\boldsymbol{A}_0^*(\mathcal{A}; P, P_0)$ in Assumption **A.5** depends on P, P_0 and on the group $\mathcal{A}$. This identification problem is related to the identification problem encountered in Independent Component Analysis (see e.g. Comon (1994), Hyvarinen et al. (2001)). To each group $\mathcal{A}$ can be attached a set $\mathcal{P}(\mathcal{A})$ such that Assumption **A.5** is satisfied for any $P, P_0 \in \mathcal{P}(\mathcal{A})$.

In order to apply the results in Eriksson and Koivunen (2004), we assume that the set of distributions $\mathcal{P}(\mathcal{A})$ includes the distributions such that the components u_{it} are independent, with different distributions and at most one Gaussian marginal distribution.

If $\mathcal{A}$ is the group of square invertible matrices, i.e. the linear group GL(n), then $\boldsymbol{A}_0^*$ is unique up to permutations and homothetic transformations of the columns (see Eriksson and Koivunen (2004), Gouriéroux et al. (2016)). From a group perspective, the set generated by the permutations and homotheties forms a sub-group M and the identifiable parameter is an element of the quotient group GL(n)$/M$.

If $\mathcal{A}$ is the group of triangular matrices with positive diagonal elements, then $\boldsymbol{A}_0^*$ is unique up to homothetic transformations of the columns. Indeed, the permutations of the columns are no longer possible. It is neither necessary to assume a priori spherical distributions, nor distributions with symmetry properties, nor to assume the existence of first and second-order moments.

We now consider examples introduced in the previous sections.

5.1.1 Newey-Steigerwald model (Example 1 continued)

Conditions for identification have been analyzed in Newey-Steigerwald (1997). As noted in their paper p. 588 (see also Th. 2): "If one additional parameter is introduced (that is in the augmented model), the identification condition for consistency is satisfied even if the symmetry condition (on the "assumed innovation" and "true innovation" densities) does not hold." In this case, the identification can be derived without introducing restrictions on the distribution of $\boldsymbol{u}$. This is compatible with the discussion in Section 5.1 of identification in the generic model. Indeed, in the one-dimensional case, the set $\mathcal{P}(\mathcal{A})$ includes Gaussian as well as non Gaussian distributions.

5.1.2 Exponential transformation model (Examples 3,4,5)

Identification can be studied directly by considering the limiting expected conditional pseudo-likelihood derived in Proposition 1 and Section 9.1 in the on-line appendix. In view of (6.2), the identification condition in the generic model is that

$$\max_{a_j} E_0 \log g \left[\exp\left(\sum_{j=1}^{J} a_j \boldsymbol{C}_j \right) \boldsymbol{u} \right] + \sum_{j=1}^{J} a_j \mathrm{Tr}(\boldsymbol{C}_j), \tag{5.1}$$

has a unique solution. Under the Eriksson and Koivunen condition on the distribution of $\boldsymbol{u}$, it is satisfied if and only if

$$\exp\left(\sum_{j=1}^{J} a_j \boldsymbol{C}_j \right) = \exp\left(\sum_{j=1}^{J} \alpha_j \boldsymbol{C}_j \right),$$

up to a permutation of columns and scale effects on the columns, implies $a_j = \alpha_j$, for $j = 1, \ldots, J$. This condition is equivalent to

$$\sum_{j=1}^{J} \beta_j \boldsymbol{C}_j = \boldsymbol{0} \quad \Rightarrow \quad \beta_j = 0, \quad j = 1, \ldots, J,$$

since matrix $\boldsymbol{0}$ is invariant by permutation of columns and scale effects on the columns. Thus, the identification in the generic model is equivalent to the linear independence of matrices $C_j, j = 1, \ldots, J$. This identification condition is clearly satisfied in Example 3 where the orthogonal projectors P_j are linearly independent. Indeed, if $\sum_{j=1}^{J} \beta_j \boldsymbol{P}_j = \boldsymbol{0}$, we also have $\sum_{j=1}^{J} \beta_j \boldsymbol{P}_j \boldsymbol{P}_k = \beta_k \boldsymbol{P}_k^2 = \beta_k \boldsymbol{P}_k = \boldsymbol{0}$, for any k, which entails $\beta_k = 0$.

The identification condition is also satisfied for any model with $J = 1$ and $\boldsymbol{C}_1 \neq \boldsymbol{0}$, as in Example 5 on volatility matrices.

5.2 Identification in the augmented model (for given P)

Assumption **A.6** concerns the identification of pseudo-true values in the augmented model. The identification depends on the properties of the generic model, but also on the specification of the index functions.

Under Assumption **A.5**, Assumption **A.6** is satisfied under the primitive condition of $\mathcal{A}$-identification defined below.

Definition 1 *Parameters* $(\boldsymbol{\theta}, \boldsymbol{\Lambda})$ *are* $\mathcal{A}$*-identified iff*

$$\boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\theta})\boldsymbol{\Lambda} = \boldsymbol{A}(\boldsymbol{x}_t, \tilde{\boldsymbol{\theta}})\tilde{\boldsymbol{\Lambda}}, \quad P_x - a.s. \quad \text{with } \boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}, \boldsymbol{\Lambda}, \tilde{\boldsymbol{\Lambda}} \in \mathcal{A},$$

implies $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$ *and* $\boldsymbol{\Lambda} = \tilde{\boldsymbol{\Lambda}}$.

In practice, two cases have to be distinguished:

- If, for a given P, $(\boldsymbol{\beta}, \boldsymbol{\Lambda})$ is identifiable in the unconstrained augmented model, we can choose $\boldsymbol{\theta} = \boldsymbol{\beta}, \boldsymbol{\Lambda}$ as new parameters, that is without changing the notation. The true distribution corresponds to the special values $\boldsymbol{\theta}_0 = \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0 = \boldsymbol{I}_n$.

- If, for a given P, $(\boldsymbol{\beta}, \boldsymbol{\Lambda})$ is not identifiable in the unconstrained augmented model, we have to reparametrize the augmented model as $\boldsymbol{y}_t = \boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\theta}(\boldsymbol{\beta}))\boldsymbol{\Lambda}\boldsymbol{u}_t$, where the function $\boldsymbol{\beta} \mapsto \boldsymbol{\theta}(\boldsymbol{\beta})$ has reduced rank, and $\boldsymbol{\Lambda}$ is still left free to vary in $\mathcal{A}$. Then, to apply Proposition 2, we have to check that this reduced rank augmented model with constrained $\boldsymbol{\theta}$ is nesting the initial model, that is, that $\boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\beta}_0)$ can be written in this new parametrization as $\overline{\boldsymbol{A}}(\boldsymbol{x}_t, \boldsymbol{\theta}(\boldsymbol{\beta}_0))\boldsymbol{\Lambda}(\boldsymbol{\beta}_0)$, say, for any $\boldsymbol{\beta}_0 \in \mathcal{B}$.

Let us now provide examples of identification analysis to show how to carefully proceed and in particular how to (re)parametrize the model of interest.

5.3 Example of identified models

In practice, the two identification issues that are the identification in the generic model and the $\mathcal{A}$-identification issue concerning the specification of the score function have to be taken into account. It is known in the standard econometric literature that the $\mathcal{A}$-identification has to be analyzed case by case according to the expression of the scores and the possible exclusion restrictions on the explanatory variables. We consider below two specific examples: we first consider a Cholesky ARCH model with linear scores, then we discuss the restrictions on the scale parameter in a model with spatial interactions.

Example 2 (cont.): Cholesky ARCH model

Let us consider a Cholesky ARCH model:

$$\boldsymbol{y}_t = \begin{pmatrix} \boldsymbol{x}_t'\boldsymbol{\beta}_{11} & 0 \\ \boldsymbol{x}_t'\boldsymbol{\beta}_{21} & \boldsymbol{x}_t'\boldsymbol{\beta}_{22} \end{pmatrix} \boldsymbol{u}_t := \boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\beta})\boldsymbol{u}_t.$$

The matrix $\boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\beta})$ belongs to the group $\mathrm{GL}(n)$ and to the group of lower triangular matrices as well. When the marginal components of $\boldsymbol{u}_t$ are independent, with different marginal distributions and at most one Gaussian distribution, we can introduce two identified augmented models that are:

$$\text{Model 1:} \qquad \boldsymbol{y}_t = \begin{pmatrix} \boldsymbol{x}_t'\boldsymbol{\theta}_{11} & \boldsymbol{x}_t'\boldsymbol{\theta}_{12} \\ \boldsymbol{x}_t'\boldsymbol{\theta}_{21} & \boldsymbol{x}_t'\boldsymbol{\theta}_{22} \end{pmatrix} \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix} \boldsymbol{u}_t,$$

for $\mathcal{A}$=GL(n), and

$$\text{Model 2:} \qquad \boldsymbol{y}_t = \begin{pmatrix} \boldsymbol{x}_t'\boldsymbol{\theta}_{11} & 0 \\ \boldsymbol{x}_t'\boldsymbol{\theta}_{21} & \boldsymbol{x}_t'\boldsymbol{\theta}_{22} \end{pmatrix} \begin{pmatrix} \lambda_{11} & 0 \\ \lambda_{21} & \lambda_{22} \end{pmatrix} \boldsymbol{u}_t,$$

for the second group. Let us assume K linearly independent explanatory variables, and denote $\boldsymbol{\Theta}_k, k = 1, \ldots, K$ the matrices of coefficients of x_{kt} in the models. A condition of $\mathcal{A}$-identification of parameters $\boldsymbol{\Theta}_k, k = 1, \ldots, K, \boldsymbol{\Lambda}$ in both augmented models is $\boldsymbol{\Theta}_1 = \boldsymbol{I}_2$. The model of interest can be written as $\boldsymbol{y}_t = \sum_{k=1}^K x_{kt}\boldsymbol{B}_k\boldsymbol{u}_t$, with $\boldsymbol{B}_k = \begin{pmatrix} \beta_{11,k} & 0 \\ \beta_{21,k} & \beta_{22,k} \end{pmatrix}$ whereas the identified augmented model is $\boldsymbol{y}_t = \sum_{k=1}^K x_{kt}\boldsymbol{\Theta}_k\Lambda\boldsymbol{u}_t$. Since $\boldsymbol{\Theta}_1 = \boldsymbol{I}_2$, we get $\boldsymbol{\Lambda}(\boldsymbol{\beta}) = \boldsymbol{B}_1$ and

$\boldsymbol{\Theta}_k(\boldsymbol{\beta}) = \boldsymbol{B}_k \boldsymbol{B}_1^{-1}, k = 2, \ldots, K$ whenever $\boldsymbol{B}_1$ is invertible. The parameters $\boldsymbol{\Theta}_k(\boldsymbol{\beta})$ are consistently estimable. Since $\boldsymbol{B}_k \boldsymbol{B}_1^{-1} = \begin{pmatrix} \frac{\beta_{11,k}}{\beta_{11,1}} & 0 \\ \frac{\beta_{21,k}}{\beta_{11,1}} - \frac{\beta_{22,k}}{\beta_{22,1}}\frac{\beta_{21,1}}{\beta_{11,1}} & \frac{\beta_{22,k}}{\beta_{22,1}} \end{pmatrix}$, the parameters in the diagonal indexes are consistently estimable up to a multiplicative factor. The transformation is more complicated for the off diagonal terms, except if one explanatory variable for instance the first one is excluded from the off-diagonal term, i.e. if $\beta_{21,1} = 0$. Then, parameters $\beta_{11,k}, \beta_{21,k}$ (resp. $\beta_{22,k}$) are identifiable up to a multiplicative factor.

Example 3 (cont.): Model with homogenous dynamic spatial interactions

Let us now discuss the identification issue for the model considered in Example 3 in Section 4.3. To understand the notion of intercept, we first need to reparametrize the generic model under the form:

$$\tilde{\boldsymbol{y}} = \exp\{-\alpha_1 \boldsymbol{P}_1 - \alpha_2(\boldsymbol{I}_n - \boldsymbol{P}_1)\}\boldsymbol{u} = \{\exp(-\alpha_1)\boldsymbol{P}_1 + \exp(-\alpha_2)(\boldsymbol{I}_n - \boldsymbol{P}_1)\}\boldsymbol{u},$$

where $\boldsymbol{P}_1 = \mathbb{1}\mathbb{1}'/n$. This is a linear transformation corresponding to a matrix with diagonal elements: $\exp(-\alpha_2) + \frac{1}{n}\{\exp(-\alpha_1) - \exp(-\alpha_2)\}$, and out-of-diagonal elements $\frac{1}{n}\{\exp(-\alpha_1) - \exp(-\alpha_2)\}$. Therefore, the parametrization a_1, a_2 of Example 3, Section 2, is such that

$$\begin{cases} a_1 &= \exp(-\alpha_2) + \frac{1}{n}\{\exp(-\alpha_1) - \exp(-\alpha_2)\}, \\ a_2 &= \frac{1}{n}\{\exp(-\alpha_1) - \exp(-\alpha_2)\} / \left[\exp(-\alpha_2) + \frac{1}{n}\{\exp(-\alpha_1) - \exp(-\alpha_2)\}\right]. \end{cases}$$

Equivalently, the parametrization in Section 4.3 is related to the one of Section 2 by:

$$\begin{cases} \alpha_1 &= -\log\{a_1(1 + (n-1)a_2)\}, \\ \alpha_2 &= -\log\{a_1(1 - a_2)\}, \end{cases}$$

which is well defined for $a_1 > 0, a_2 \in \left(-\frac{1}{n-1}, 1\right)$ (see Section 2). As seen in Section 4.3, the intercepts are parameters μ_1, μ_2, say, to be added to the $\alpha_i(\boldsymbol{x}_t, \boldsymbol{\beta})$'s, $i = 1, 2$, in the econometric model. Therefore, an identification problem arises in the initial parametrization $a_{1t} = a_1(\boldsymbol{x}_t, \boldsymbol{\beta})$, $a_{2t} = a_2(\boldsymbol{x}_t, \boldsymbol{\beta})$ if and only if there is a scale parameter in either $a_{1t}(1 + (n-1)a_{2t})$, or $a_{1t}(1 - a_{2t})$. This occurs if there is a scale parameter in a_{1t}, or a scale parameter in $1 - a_{2t}$, or a scale parameter in $1 + (n-1)a_{2t}$. Such scale parameters are constrained by the conditions on the admissible values of a_{1t}, a_{2t}. For instance if $0 < a_{2t} < 1$ for all t, the associated scale parameter has to belong to $(0, 1)$.

6 Asymptotic distribution of the PML estimator

The asymptotic theory for PML estimators was initially developed in the i.i.d. setting (see e.g. White (1994), Gouriéroux, Monfort (1995), chap. 24). In our framework, we will assume independence between $\boldsymbol{u}_t$ and the past of $\boldsymbol{x}_t$, a stronger condition than **A.3**:

Assumption A.8: $\boldsymbol{u}_t$ is independent from the $\boldsymbol{x}_{t-i}$, for $i \geq 0$.

Under the Assumptions **A.1-A.8** and other regularity conditions (see the on-line Appendix), the PML estimator is asymptotically normal with asymptotic variance-covariance matrix obtained by a sandwich formula:

$$V_{as}\left[\sqrt{T}\begin{pmatrix}\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_0\\ \hat{\boldsymbol{\lambda}}_T-\boldsymbol{\lambda}_0^*\end{pmatrix}\right]=\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1},$$

where

$$\boldsymbol{A}=E_0\left[-\frac{\partial^2}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}'}l_t(\boldsymbol{\theta}_0,\boldsymbol{\lambda}_0^*)\right],\quad \boldsymbol{B}=V_0\left[\frac{\partial}{\partial\boldsymbol{\vartheta}}l_t(\boldsymbol{\theta}_0,\boldsymbol{\lambda}_0^*)\right],$$

$\boldsymbol{\vartheta}=(\boldsymbol{\theta}',\boldsymbol{\lambda}')'=(\vartheta_j)_{1\leq j\leq K+J}$ and $l_t(\boldsymbol{\vartheta})=l[a(\boldsymbol{x}_t;\boldsymbol{\theta})*\boldsymbol{\lambda},\boldsymbol{y}_t]$.

The expression of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ simplifies when the group is commutative[9], as in the exponential transformation model of Section 4.3, namely:

$$\boldsymbol{y}_t=\exp\left(-\sum_{j=1}^{J}[a_j(\boldsymbol{x}_t,\boldsymbol{\theta}_0)+\lambda_{0j}]\boldsymbol{C}_j\right)\boldsymbol{u}_t, \tag{6.1}$$

where the $n\times n$ matrices $\boldsymbol{C}_j, j=1,\ldots,J$ commute. The limiting pseudo log-likelihood in the generic model without explanatory variable is:

$$L(\boldsymbol{a};P,P_0) = E_0 l(\boldsymbol{u},\boldsymbol{a}),\quad \text{where}\quad l(\boldsymbol{u},\boldsymbol{a})=\log g\left[\exp\left(\sum_{j=1}^{J}a_j\boldsymbol{C}_j\right)\boldsymbol{u}\right]+\sum_{j=1}^{J}a_j\mathrm{Tr}(\boldsymbol{C}_j), \tag{6.2}$$

and the pseudo log-likelihood corresponding to date t in Model (6.1) is $l(\boldsymbol{y}_t,\boldsymbol{a}(\boldsymbol{x}_t,\boldsymbol{\theta})+\boldsymbol{\lambda})$.

The asymptotic distribution of the PML estimator will follow from the property of martingale difference of the pseudo score, and the asymptotic variance-covariance matrix of the PML estimator is obtained by a simplified sandwich formula.

Proposition 3 *Under the Assumptions* **A.1-A.8** *and other regularity conditions (see the on-line Appendix), the PML estimator is asymptotically normal:*

$$\sqrt{T}\begin{pmatrix}\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_0\\ \\ \hat{\boldsymbol{\lambda}}_T-\boldsymbol{\lambda}_0^*\end{pmatrix}\xrightarrow{d}N\left(\boldsymbol{0},\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}\right),$$

where $\boldsymbol{\lambda}_0^*=\boldsymbol{\lambda}_0+\boldsymbol{a}_0^*$ *and:*

$$\boldsymbol{B}=E_x\left[\begin{pmatrix}\frac{\partial\boldsymbol{a}'(\boldsymbol{x}_t;\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\\ \boldsymbol{I}_J\end{pmatrix}\boldsymbol{K}\begin{pmatrix}\frac{\partial\boldsymbol{a}'(\boldsymbol{x}_t;\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\\ \boldsymbol{I}_J\end{pmatrix}'\right],\qquad \boldsymbol{A}=E_x\left[\begin{pmatrix}\frac{\partial\boldsymbol{a}'(\boldsymbol{x}_t;\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\\ \boldsymbol{I}_J\end{pmatrix}\boldsymbol{L}\begin{pmatrix}\frac{\partial\boldsymbol{a}'(\boldsymbol{x}_t;\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\\ \boldsymbol{I}_J\end{pmatrix}'\right],$$

where the $J\times J$ *matrices* $\boldsymbol{K}$ *and* $\boldsymbol{L}$ *only depend on the true and pseudo densities of* $\boldsymbol{u}_t$ *and are displayed in the proof.*

Proof: See the on-line Appendix.

[9] When the group is not commutative, for example for the Cholesky ARCH model, this simplification does not exist.

7 Concluding Remarks

We have provided group transformation models such that the PML estimators of functions $\boldsymbol{\theta}(\boldsymbol{\beta}_0)$ of the parameter of interest are consistent for any "regular" pseudo-distribution, when the error $\boldsymbol{u}_t$ is independent from the explanatory variable $\boldsymbol{x}_t$. For a given transformation $\boldsymbol{c}$ and a given index function $\boldsymbol{a}(\boldsymbol{x}_t, \boldsymbol{\theta})$, we get an infinite number of PML estimators $\boldsymbol{\theta}$. We have shown that for multivariate models based on exponential transformations, under appropriate regularity conditions, these PML estimators are asymptotically normal with asymptotic variance-covariance matrices obtained by the so-called sandwich formula. These matrices can be used to select an accurate PML among all the consistent ones and to construct misspecification tests. The existence of representation (3.10), with identifiable parameter $\boldsymbol{\theta}_0$ and estimable errors distribution, suggests studying adaptive estimation methods, as in Hafner and Rombouts (2007) for the Gaussian PML in multivariate volatility models, in which an estimator of the errors distribution is used to efficiently estimate $\boldsymbol{\theta}_0$. This is left for further research.

References

[1] Anderson, T.W. (1970): "Estimation of Covariance Matrices which are Linear Combinations or Whose Inverse are Linear Combinations of Given Matrices", in Essays in Probability and Statistics, 1-24, University of North-Carolina Press, Chapel Hill.

[2] Arsigny, V., Fillard, P., Pennec, X., and N., Ayache (2007): "Geometric Means in a Novel Vector Space Structure on Symmetric Positive-Definite Matrices", Siam J. Matrix Anal. Appl., 29, 328-347.

[3] Banerjee, S., Carlin, B., and A., Gelfand (2004): "Hierarchical Modelling and Analysis for Spatial Data", Chapman Hall.

[4] Bollerslev, T., and J., Woolridge (1992): "Quasi-Maximum Likelihood Estimators and Inference in Dynamic Models with Time Varying Covariance", Econometric Reviews, 11, 143-172.

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

[6] Cressie, N. (1993): "Statistics for Spatial Data", New-York, Wiley.

[7] Dellaportas, P., and M., Pourahmadi (2012): "Cholesky-GARCH Models with Application to Finance", Stat. Comput., 22, 849-855.

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

[9] Fan, J., Qi, L. and D., Xiu (2014): "Quasi Maximum Likelihood Estimation of GARCH Models with Heavy-Tailed Likelihoods", Journal of Business and Economic Statistics, 32, 193-198.

[10] Felsenstein, J. (1973): "Maximum-Likelihood Estimation of Evolutionary Trees from Continuous Characteristics", American Journal of Human Genetics, 25, 471-492.

[11] Fiorentini, G., and E., Sentana (2016): "Consistent non-Gaussian Pseudo Maximum Likelihood Estimators", DP CEMFI.

[12] Francq, C., Lepage, G. and J-M., Zakoïan (2011): "Two-Stage non Gaussian QML Estimation of GARCH models and Testing the Efficiency of the Gaussian QMLE", Journal of Econometrics, 165, 246-257.

[13] Goldsmith-Pinkham, P., and G., Imbens (2013): "Social Networks and the Identification of Peer Effects", Journal of Business and Economics Statistics, 31, 253-264.

[14] Gouriéroux, C., and A., Monfort (1995) : "Statistics and Econometric Models", Cambridge University Press.

[15] Gouriéroux, C., Monfort, A., and J.P., Renne (2016): "Statistical Inference for Independent Component Analysis", Journal of Econometrics, 196, 111-126.

[16] Gouriéroux, C., Monfort, A., and A., Trognon (1984): "Pseudo Maximum Likelihood Methods: Theory", Econometrica, 52, 681-700.

[17] Hafner, C.M. and J.V.K. Rombouts (2007): "Semiparametric multivariate volatility models", Econometric Theory, 23, 251-280.

[18] Huang, Z., Wang, R., Shao, S., Li, X., and X., Chen (2015): "Log Euclidean Metric Learning on Symmetric Positive Definite Manifolds with Applications to Image Set Classification", in Proceedings of the 32nd International Conference on Machine Learning, JMLR, vol 37.

[19] Hyvarinen, A., Kaihunen, J., and E. Oja (2001): "Independent Component Analysis", Nee-York, Wiley.

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

[21] Newey, W., and D., Steigerwald (1997): "Asymptotic Bias for Quasi-Maximum Likelihood Estimators in Conditional Heteroskedastic Models", Econometrica, 65, 587-599.

[22] White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models", Econometrica, 50, 1-26.

[23] White, H. (1994): "Estimation, Inference and Specification Analysis", Cambridge Univ. Press.

[24] Yuan, Y., Zhu, H., Lin, W., and J., Marron (2012): "Local Polynomial Regression for Symmetric Positive Definite Matrices", Journal of the Royal Statistical Society, Series B, 74, 697-719.

[25] Zwiernik, P., Uhler, C., and D., Richards (2016): "Maximum Likelihood Estimation for Linear Gaussian Covariance Models", JRSS B, 1-24.

Consistent Pseudo-Maximum Likelihood Estimators and Groups of Transformations:
On-line Appendix (not for publication)

This document consists of two sections of additional results: i) Regularity conditions for Proposition 3 and sketch of proof; ii) Derivatives of functions based on exponential of matrices.

8 Regularity conditions for Proposition 3

Let $\boldsymbol{\vartheta} = (\boldsymbol{\theta}', \boldsymbol{\lambda}')' = (\vartheta_j)_{1\leq j\leq K+J}$ and $l_t(\boldsymbol{\vartheta}) = l[a(\boldsymbol{x}_t;\boldsymbol{\theta}) + \boldsymbol{\lambda}, \boldsymbol{y}_t]$.

Together with Assumptions **A.1-A.7**, Proposition 3 requires the following assumptions:

Assumption A.9: $\boldsymbol{\vartheta}_0^* = (\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*)$ belongs to the interior of $\boldsymbol{\theta}(\mathcal{B}) \times \mathcal{A}$.

Assumption A.10: For any $\boldsymbol{x}$, the function $\boldsymbol{\theta} \to a(\boldsymbol{x};\boldsymbol{\theta})$ has continuous third-order derivatives. The pseudo-density function g is three times continuously differentiable.

Assumption A.11: The matrices $\boldsymbol{K}$ and $\boldsymbol{L}$ defined in the proof are positive definite.

Assumption A.12: For at least one $j \in \{1,\dots,J\}$, the matrix $\mathrm{V}_0\left(\frac{\partial a_j}{\partial \boldsymbol{\theta}}(x_t, \boldsymbol{\theta}_0)\right)$ is positive definite.

Assumption A.13: There exists a neighborhood $V(\boldsymbol{\vartheta}_0^*)$ of $\boldsymbol{\vartheta}_0^*$ such that, for $i, j = 1, \dots, r$,

for all $\boldsymbol{\vartheta} \in V(\boldsymbol{\vartheta}_0^*)$, the process $\left\{\frac{\partial}{\partial \boldsymbol{\vartheta}'}\left(\frac{\partial^2}{\partial\vartheta_i\partial\vartheta_j}\ell_t(\boldsymbol{\vartheta})\right)\right\}$ is strictly stationary and ergodic, and,

$$E_0 \sup_{\boldsymbol{\vartheta}\in V(\boldsymbol{\vartheta}_0^*)} \left\|\frac{\partial}{\partial \boldsymbol{\vartheta}'}\left(\frac{\partial^2}{\partial\vartheta_i\partial\vartheta_j}\ell_t(\boldsymbol{\vartheta})\right)\right\| < \infty.$$

9 Proof of Proposition 3

In this Section, we will explain how to use an appropriate Central Limit Theorem (CLT), and we will derive the asymptotic covariance matrix.

9.1 The pseudo-score

For Model (6.1), the pseudo log-likelihood for one observation takes the form:

$$l_t(\boldsymbol{\vartheta}) = l_t(\boldsymbol{\theta}, \boldsymbol{\lambda}) \quad = \quad \log g\left[\exp\left\{\sum_{j=1}^{J}[a_j(\boldsymbol{x}_t;\boldsymbol{\theta}) + \lambda_j]\boldsymbol{C}_j\right\}\boldsymbol{y}_t\right] + \sum_{j=1}^{J}[a_j(\boldsymbol{x}_t;\boldsymbol{\theta}) + \lambda_j]\mathrm{Tr}(\boldsymbol{C}_j).$$

Let

$$z_t(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \exp\left\{\sum_{j=1}^{J}[a_j(\boldsymbol{x}_t;\boldsymbol{\theta}) + \lambda_j]\boldsymbol{C}_j\right\}\boldsymbol{y}_t.$$

For $\boldsymbol{\gamma} = (\mathrm{Tr}(\boldsymbol{C}_1), \ldots, \mathrm{Tr}(\boldsymbol{C}_J))'$, we have:

$$l_t(\boldsymbol{\theta}, \boldsymbol{\lambda}) \quad = \quad \log g\{z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})\} + \boldsymbol{\gamma}'\{\boldsymbol{a}(\boldsymbol{x}_t; \boldsymbol{\theta}) + \boldsymbol{\lambda}\}.$$

Using the computations of Section 10.3, it follows that:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}'} l_t(\boldsymbol{\theta}, \boldsymbol{\lambda}) &= \frac{\partial \log g}{\partial \boldsymbol{u}'} \{z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})\} \frac{\partial z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\theta}'} + \boldsymbol{\gamma}'\{\boldsymbol{a}(\boldsymbol{x}_t; \boldsymbol{\theta}) + \boldsymbol{\lambda}\}, \\
&= \left(\frac{\partial \log g}{\partial \boldsymbol{u}'} \{z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})\} \boldsymbol{C}[\boldsymbol{I}_J \otimes z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})] + \boldsymbol{\gamma}' \right) \frac{\partial \boldsymbol{a}(\boldsymbol{x}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \\
&:= h_t'(\boldsymbol{\theta}, \boldsymbol{\lambda}) \frac{\partial \boldsymbol{a}(\boldsymbol{x}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}.
\end{aligned}$$

Proceeding similarly with parameter $\boldsymbol{\lambda}$, we find that:

$$\frac{\partial}{\partial \boldsymbol{\vartheta}} l_t(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \begin{pmatrix} \frac{\partial}{\partial \boldsymbol{\theta}} l_t(\boldsymbol{\theta}, \boldsymbol{\lambda}) \\ \frac{\partial}{\partial \boldsymbol{\lambda}} l_t(\boldsymbol{\theta}, \boldsymbol{\lambda}) \end{pmatrix} \quad = \quad \begin{pmatrix} \frac{\partial \boldsymbol{a}'(\boldsymbol{x}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ \boldsymbol{I}_J \end{pmatrix} h_t(\boldsymbol{\theta}, \boldsymbol{\lambda}).$$

9.2 The martingale difference property

Replacing $\boldsymbol{y}_t$ by $\exp\left\{-\sum_{j=1}^{J}[a_j(\boldsymbol{x}_t; \boldsymbol{\theta}_0) + \lambda_{0j}]\boldsymbol{C}_j\right\} \boldsymbol{u}_t$, we find that

$$z_t(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*) = \exp\left\{\sum_{j=1}^{J} (\lambda_{0j}^* - \lambda_{0j}) \boldsymbol{C}_j\right\} \boldsymbol{u}_t := \boldsymbol{\Gamma}(\boldsymbol{\lambda}_0^* - \boldsymbol{\lambda}_0)\boldsymbol{u}_t.$$

Thus,

$$\frac{\partial}{\partial \boldsymbol{\vartheta}} l_t(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*) = \begin{pmatrix} \frac{\partial}{\partial \boldsymbol{\theta}} l_t(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*) \\ \frac{\partial}{\partial \boldsymbol{\lambda}} l_t(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*) \end{pmatrix} \quad = \quad \begin{pmatrix} \frac{\partial a'(\boldsymbol{x}_t; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \\ \boldsymbol{I}_J \end{pmatrix} \boldsymbol{k}(\boldsymbol{u}_t),$$

where

$$\boldsymbol{k}(\boldsymbol{u}_t) = [\boldsymbol{I}_J \otimes (\boldsymbol{\Gamma}(\boldsymbol{\lambda}_0^* - \boldsymbol{\lambda}_0)\boldsymbol{u}_t)']\boldsymbol{C}' \frac{\partial \log g}{\partial \boldsymbol{u}} \{\boldsymbol{\Gamma}(\boldsymbol{\lambda}_0^* - \boldsymbol{\lambda}_0)\boldsymbol{u}_t\} + \boldsymbol{\gamma}.$$

Noting that

$$\boldsymbol{k}(\boldsymbol{u}_t) = [\boldsymbol{I}_J \otimes (\boldsymbol{\Gamma}(\boldsymbol{a}_0^*)\boldsymbol{u}_t)']\boldsymbol{C}' \frac{\partial \log g}{\partial \boldsymbol{u}} \{\boldsymbol{\Gamma}(\boldsymbol{a}_0^*)\boldsymbol{u}_t\} + \boldsymbol{\gamma},$$

where $\boldsymbol{a}_0^*$ is defined in **A.5**, we have $E\{\boldsymbol{k}(\boldsymbol{u}_t)\} = 0$ from the first-order conditions in the generic model. Thus, using Assumption **A.8**, $\left(\frac{\partial}{\partial \boldsymbol{\vartheta}} l_t(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*), \mathcal{F}_t\right)$ is a martingale difference sequence, where $\mathcal{F}_t = \sigma\{(\boldsymbol{u}_i, \boldsymbol{x}_i), i \leq t\}$. The asymptotic normality follows from applying a CLT for the square integrable, ergodic and stationary martingale difference (see Billingsley, 1961). We get

$$V_{as}\left[\sqrt{T}\begin{pmatrix} \hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\lambda}}_T - \boldsymbol{\lambda}_0^* \end{pmatrix}\right] = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1},$$

where

$$\boldsymbol{A} = E_0\left[-\frac{\partial^2}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}'} l_t(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*)\right], \quad \boldsymbol{B} = V_0\left[\frac{\partial}{\partial \boldsymbol{\vartheta}} l_t(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*)\right].$$

9.3 Computation of matrix the asymptotic covariance matrix

We have

$$\boldsymbol{B} = V_0\left[\frac{\partial}{\partial \boldsymbol{\vartheta}} l_t(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*)\right] \quad = \quad E_0\left[\begin{pmatrix} \frac{\partial a'(\boldsymbol{x}_t;\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \\ \boldsymbol{I}_J \end{pmatrix} \boldsymbol{K} \begin{pmatrix} \frac{\partial a'(\boldsymbol{x}_t;\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \\ \boldsymbol{I}_J \end{pmatrix}'\right], \qquad \boldsymbol{K} = V_0\left[\boldsymbol{k}(\boldsymbol{u}_t)\right].$$

Now,

$$\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} l_t(\boldsymbol{\theta}, \boldsymbol{\lambda}) \quad = \quad \frac{\partial h_t'(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{a}(\boldsymbol{x}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} + \{h_t'(\boldsymbol{\theta}, \boldsymbol{\lambda}) \otimes \boldsymbol{I}_p\} A(\boldsymbol{x}_t; \boldsymbol{\theta}),$$

where $A(\boldsymbol{x}_t; \boldsymbol{\theta})$ is the $Jp \times p$ matrix:

$$A(\boldsymbol{x}_t; \boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial^2 a_1(\boldsymbol{x}_t;\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \\ \vdots \\ \frac{\partial^2 a_J(\boldsymbol{x}_t;\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \end{pmatrix}.$$

Noting that:

$$h_t'(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \left[\frac{\partial \log g}{\partial \boldsymbol{u}'}\{z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})\} \boldsymbol{C}_1 z_t(\boldsymbol{\theta}, \boldsymbol{\lambda}), \ldots, \frac{\partial \log g}{\partial \boldsymbol{u}'}\{z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})\} \boldsymbol{C}_J z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})\right] + \boldsymbol{\gamma}',$$

we compute:

$$\begin{aligned} & \frac{\partial}{\partial \boldsymbol{\theta}}\left\{\frac{\partial \log g}{\partial \boldsymbol{u}'}\{z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})\} \boldsymbol{C}_j z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})\right\} \\ = \; & \frac{\partial}{\partial \boldsymbol{\theta}}\left\{\frac{\partial \log g}{\partial \boldsymbol{u}'}\{z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})\}\right\} \boldsymbol{C}_j z_t(\boldsymbol{\theta}, \boldsymbol{\lambda}) + \left[\frac{\partial}{\partial \boldsymbol{\theta}}\{\boldsymbol{C}_j z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})\}'\right]\left\{\frac{\partial \log g}{\partial \boldsymbol{u}'}\{z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})\}\right\}' \\ = \; & \frac{\partial z_t'(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\theta}} \frac{\partial^2 \log g}{\partial \boldsymbol{u} \partial \boldsymbol{u}'}\{z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})\} \boldsymbol{C}_j z_t(\boldsymbol{\theta}, \boldsymbol{\lambda}) \\ & + \frac{\partial \boldsymbol{a}'(\boldsymbol{x}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}[\boldsymbol{I}_J \otimes z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})]' \boldsymbol{C}' \boldsymbol{C}_j' \frac{\partial \log g}{\partial \boldsymbol{u}}\{z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})\} \\ = \; & \frac{\partial \boldsymbol{a}'(\boldsymbol{x}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}[\boldsymbol{I}_J \otimes z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})]' \boldsymbol{C}' \\ & \times \left\{\frac{\partial^2 \log g}{\partial \boldsymbol{u} \partial \boldsymbol{u}'}\{z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})\} \boldsymbol{C}_j z_t(\boldsymbol{\theta}, \boldsymbol{\lambda}) + \boldsymbol{C}_j' \frac{\partial \log g}{\partial \boldsymbol{u}}\{z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})\}\right\}. \end{aligned}$$

It follows that:

$$\begin{aligned} & \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} l_t(\boldsymbol{\theta}, \boldsymbol{\lambda}) \\ = \; & \frac{\partial \boldsymbol{a}'(\boldsymbol{x}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}[\boldsymbol{I}_J \otimes z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})]' \boldsymbol{C}' \\ & \times \sum_{j=1}^{J}\left\{\frac{\partial^2 \log g}{\partial \boldsymbol{u} \partial \boldsymbol{u}'}\{z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})\} \boldsymbol{C}_j z_t(\boldsymbol{\theta}, \boldsymbol{\lambda}) + \boldsymbol{C}_j' \frac{\partial \log g}{\partial \boldsymbol{u}}\{z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})\}\right\} \frac{\partial a_j(\boldsymbol{x}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \\ & + \{h_t'(\boldsymbol{\theta}, \boldsymbol{\lambda}) \otimes \boldsymbol{I}_p\} A(\boldsymbol{x}_t; \boldsymbol{\theta}) \\ = \; & \frac{\partial \boldsymbol{a}'(\boldsymbol{x}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}[\boldsymbol{I}_J \otimes z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})]' \boldsymbol{C}' \frac{\partial^2 \log g}{\partial \boldsymbol{u} \partial \boldsymbol{u}'}\{z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})\} \boldsymbol{C}[\boldsymbol{I}_J \otimes z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})] \frac{\partial \boldsymbol{a}(\boldsymbol{x}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \\ & + \frac{\partial \boldsymbol{a}'(\boldsymbol{x}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}[\boldsymbol{I}_J \otimes z_t(\boldsymbol{\theta}, \boldsymbol{\lambda})]' \boldsymbol{C}' \boldsymbol{G}[\boldsymbol{z}_t(\boldsymbol{\theta}, \boldsymbol{\lambda})] \frac{\partial \boldsymbol{a}(\boldsymbol{x}_t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} + \{h_t'(\boldsymbol{\theta}, \boldsymbol{\lambda}) \otimes \boldsymbol{I}_p\} A(\boldsymbol{x}_t; \boldsymbol{\theta}), \end{aligned}$$

where $G(\boldsymbol{u})$ is the $n \times J$ matrix:

$$\boldsymbol{G}(\boldsymbol{u}) = \left[\boldsymbol{C}_1' \frac{\partial \log g}{\partial \boldsymbol{u}}(\boldsymbol{u}) \;\; \boldsymbol{C}_2' \frac{\partial \log g}{\partial \boldsymbol{u}}(u) \;\; \ldots \;\; \boldsymbol{C}_J' \frac{\partial \log g}{\partial \boldsymbol{u}}(\boldsymbol{u})\right].$$

Note that the first-order conditions imply $Eh_t'(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*) = 0$. Therefore

$$\boldsymbol{A} = E_0\left[-\frac{\partial^2}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}'} l_t(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0^*)\right] \;\; = \;\; E_0\left[\begin{pmatrix} \frac{\partial a'(\boldsymbol{x}_t;\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \\ \boldsymbol{I}_J \end{pmatrix} \boldsymbol{L} \begin{pmatrix} \frac{\partial a'(\boldsymbol{x}_t;\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \\ \boldsymbol{I}_J \end{pmatrix}'\right]$$

where:

$$\begin{aligned} \boldsymbol{L} \;\; = \;\; & -E_0\Big\{[\boldsymbol{I}_J \otimes \boldsymbol{\Gamma}(\boldsymbol{\lambda}_0^* - \boldsymbol{\lambda}_0)\boldsymbol{u}_t]'\boldsymbol{C}' \frac{\partial^2 \log g}{\partial \boldsymbol{u} \partial \boldsymbol{u}'} \{\boldsymbol{\Gamma}(\boldsymbol{\lambda}_0^* - \boldsymbol{\lambda}_0)\boldsymbol{u}_t\} \boldsymbol{C}[\boldsymbol{I}_J \otimes \boldsymbol{\Gamma}(\boldsymbol{\lambda}_0^* - \boldsymbol{\lambda}_0)\boldsymbol{u}_t] \\ & +[\boldsymbol{I}_J \otimes \boldsymbol{\Gamma}(\boldsymbol{\lambda}_0^* - \boldsymbol{\lambda}_0)\boldsymbol{u}_t]'\boldsymbol{C}'\boldsymbol{G}\left(\boldsymbol{\Gamma}(\boldsymbol{\lambda}_0^* - \boldsymbol{\lambda}_0)\boldsymbol{u}_t\right)\Big\}. \end{aligned}$$

10 Derivatives of functions based on exponential of matrices

10.1 Derivatives of $a \to \log g(e^{a\boldsymbol{C}}\boldsymbol{y})$

For $a \in \mathbb{R}$, $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{C}$ a $n \times n$ matrix, $g : \mathbb{R}^n \to \mathbb{R}^+$ a function,

$$\begin{aligned} \frac{\partial}{\partial a}(e^{a\boldsymbol{C}}\boldsymbol{y}) \;\; &= \;\; \boldsymbol{C}e^{a\boldsymbol{C}}\boldsymbol{y}, \qquad \frac{\partial^2}{\partial a^2}(e^{a\boldsymbol{C}}\boldsymbol{y}) = \boldsymbol{C}^2 e^{a\boldsymbol{C}}\boldsymbol{y} \\ \frac{\partial}{\partial a} \log g(e^{a\boldsymbol{C}}\boldsymbol{y}) \;\; &= \;\; \left[\frac{\partial \log g}{\partial \boldsymbol{u}'}(e^{a\boldsymbol{C}}\boldsymbol{y})\right] \boldsymbol{C}e^{a\boldsymbol{C}}\boldsymbol{y}, \\ \frac{\partial^2}{\partial a^2} \log g(e^{a\boldsymbol{C}}\boldsymbol{y}) \;\; &= \;\; \left(\boldsymbol{C}e^{a\boldsymbol{C}}\boldsymbol{y}\right)' \left[\frac{\partial^2 \log g}{\partial \boldsymbol{u} \partial \boldsymbol{u}'}(e^{a\boldsymbol{C}}\boldsymbol{y})\right] \boldsymbol{C}e^{a\boldsymbol{C}}\boldsymbol{y} + \left(\boldsymbol{C}^2 e^{a\boldsymbol{C}}\boldsymbol{y}\right)' \left[\frac{\partial \log g}{\partial \boldsymbol{u}}(e^{a\boldsymbol{C}}\boldsymbol{y})\right]. \end{aligned}$$

10.2 Derivatives of $\boldsymbol{\theta} \to e^{a(\boldsymbol{\theta})\boldsymbol{C}}\boldsymbol{y}$ and $\boldsymbol{\theta} \to \log g(e^{a(\boldsymbol{\theta})\boldsymbol{C}}\boldsymbol{y})$

For $a : \mathbb{R}^p \to \mathbb{R}$ and $\boldsymbol{\theta} \in \mathbb{R}^p$,

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}'}\left\{e^{a(\boldsymbol{\theta})\boldsymbol{C}}\boldsymbol{y}\right\} \;\; &= \;\; \boldsymbol{C}e^{a(\boldsymbol{\theta})\boldsymbol{C}}\boldsymbol{y}\frac{\partial a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}, \\ \frac{\partial}{\partial \boldsymbol{\theta}} \log g\left\{e^{a(\boldsymbol{\theta})\boldsymbol{C}}\boldsymbol{y}\right\} \;\; &= \;\; \frac{\partial}{\partial a} \log g\left\{e^{a(\boldsymbol{\theta})\boldsymbol{C}}\boldsymbol{y}\right\} \cdot \frac{\partial a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= \;\; \left\{\left[\frac{\partial \log g}{\partial \boldsymbol{u}'}\left\{e^{a(\boldsymbol{\theta})\boldsymbol{C}}\boldsymbol{y}\right\}\right] \boldsymbol{C}e^{a(\boldsymbol{\theta})\boldsymbol{C}}\boldsymbol{y}\right\} \cdot \frac{\partial a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \\ \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log g\left\{e^{a(\boldsymbol{\theta})\boldsymbol{C}}\boldsymbol{y}\right\} \;\; &= \;\; \frac{\partial}{\partial a} \log g\left\{e^{a(\boldsymbol{\theta})\boldsymbol{C}}\boldsymbol{y}\right\} \cdot \frac{\partial^2 a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} + \frac{\partial^2}{\partial a^2} \log g(e^{a\boldsymbol{C}}\boldsymbol{y}) \cdot \frac{\partial a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \\ &= \;\; \left\{\left[\frac{\partial \log g}{\partial \boldsymbol{u}'}\left\{e^{a(\boldsymbol{\theta})\boldsymbol{C}}\boldsymbol{y}\right\}\right] \boldsymbol{C}e^{a(\boldsymbol{\theta})\boldsymbol{C}}\boldsymbol{y}\right\} \cdot \frac{\partial^2 a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} + \\ & \quad \left\{\left(\boldsymbol{C}e^{a(\boldsymbol{\theta})\boldsymbol{C}}\boldsymbol{y}\right)' \left[\frac{\partial^2 \log g}{\partial \boldsymbol{u} \partial \boldsymbol{u}'}(e^{a(\boldsymbol{\theta})\boldsymbol{C}}\boldsymbol{y})\right] \boldsymbol{C}e^{a(\boldsymbol{\theta})\boldsymbol{C}}\boldsymbol{y}\right. \\ & \qquad \left. + \left(\boldsymbol{C}^2 e^{a(\boldsymbol{\theta})\boldsymbol{C}}\boldsymbol{y}\right)' \left[\frac{\partial \log g}{\partial \boldsymbol{u}}(e^{a(\boldsymbol{\theta})\boldsymbol{C}}\boldsymbol{y})\right]\right\} \cdot \frac{\partial a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial a(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}. \end{aligned}$$

In these equalities, "$\cdot$" indicates the multiplication of a matrix by a scalar.

10.3 Derivatives of $\boldsymbol{\theta} \to \exp\left\{\sum_{j=1}^{J}[a_j(\boldsymbol{x}_t;\boldsymbol{\theta})+\lambda_j]\boldsymbol{C}_j\right\}\boldsymbol{y}_t$

Let $\boldsymbol{z}_t(\lambda,\boldsymbol{\theta}) = \exp\left\{\sum_{j=1}^{J}[a_j(\boldsymbol{x}_t;\boldsymbol{\theta})+\lambda_j]\boldsymbol{C}_j\right\}\boldsymbol{y}_t$, where $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_J)' \in \mathbb{R}^J$, $a_j(\cdot)$ are real valued functions with $\boldsymbol{\theta} \in \mathbb{R}^p$, $\boldsymbol{y}_t \in \mathbb{R}^n$, $\boldsymbol{C}_j$ are $n \times n$ matrices. Let $a(\boldsymbol{x}_t;\boldsymbol{\theta}) = (a_1(\boldsymbol{x}_t;\boldsymbol{\theta}),\ldots,a_J(\boldsymbol{x}_t;\boldsymbol{\theta}))'$. For $i = 1,\ldots,J$, let the $n \times 1$ vectors

$$\begin{aligned}\boldsymbol{z}_t^{(i)}(\boldsymbol{\theta},\boldsymbol{\lambda}) &= \exp\left\{\sum_{j=1}^{i-1}[a_j(\boldsymbol{x}_t;\boldsymbol{\theta})+\lambda_j]\boldsymbol{C}_j\right\}\boldsymbol{C}_i \exp\left\{[a_i(\boldsymbol{x}_t;\boldsymbol{\theta})+\lambda_i]\boldsymbol{C}_i\right\} \\ &\quad \times \exp\left\{\sum_{j=i+1}^{J}[a_j(\boldsymbol{x}_t;\boldsymbol{\theta})+\lambda_j]\boldsymbol{C}_j\right\}\boldsymbol{y}_t,\end{aligned}$$

where the first and last sums are replaced by 0 when $i = 1$ and $i = J$, respectively. Let the $n \times J$ block-matrix

$$\boldsymbol{Z}_t(\boldsymbol{\theta},\boldsymbol{\lambda}) = [\boldsymbol{z}_t^{(1)}(\boldsymbol{\theta},\boldsymbol{\lambda})|\ldots|\boldsymbol{z}_t^{(J)}(\boldsymbol{\theta},\boldsymbol{\lambda})].$$

We have

$$\frac{\partial \boldsymbol{z}_t(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial \boldsymbol{\theta}'} = \boldsymbol{Z}_t(\boldsymbol{\theta},\boldsymbol{\lambda})\frac{\partial a(\boldsymbol{x}_t;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}.$$

When the matrices $\boldsymbol{C}_j$ commute, we have $\boldsymbol{z}_t^{(i)}(\boldsymbol{\theta},\boldsymbol{\lambda}) = \boldsymbol{C}_i\boldsymbol{z}_t(\boldsymbol{\theta},\boldsymbol{\lambda})$ and

$$\boldsymbol{Z}_t(\boldsymbol{\theta},\boldsymbol{\lambda}) = [\boldsymbol{C}_1\boldsymbol{z}_t(\boldsymbol{\theta},\boldsymbol{\lambda})|\ldots|\boldsymbol{C}_J\boldsymbol{z}_t(\boldsymbol{\theta},\boldsymbol{\lambda})] = \boldsymbol{C}[\boldsymbol{I}_J \otimes \boldsymbol{z}_t(\boldsymbol{\theta},\boldsymbol{\lambda})],$$

where $\boldsymbol{C} = [\boldsymbol{C}_1|\ldots|\boldsymbol{C}_J]$. Thus, when the $\boldsymbol{C}_j$ commute,

$$\frac{\partial \boldsymbol{z}_t(\boldsymbol{\theta},\boldsymbol{\lambda})}{\partial \boldsymbol{\theta}'} = \boldsymbol{C}[\boldsymbol{I}_J \otimes \boldsymbol{z}_t(\boldsymbol{\theta},\boldsymbol{\lambda})]\frac{\partial a(\boldsymbol{x}_t;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}.$$

References

[1] Billingsley, P.(1961): "The Lindeberg-Lévy Theorem for Martingales", *Proceedings of the American Mathematical Society* 12, 788–792.