Non-Independent Components Analysis

BSE Working Paper 1358
September 2022 (Revised: October 2023)

Geert Mesters, Piotr Zwiernik

bse.eu/research

NON-INDEPENDENT COMPONENTS ANALYSIS

GEERT MESTERS AND PIOTR ZWIERNIK

ABSTRACT. A seminal result in the ICA literature states that for $AY = \varepsilon$, if the components of ε are independent and at most one is Gaussian, then A is identified up to sign and permutation of its rows [Comon, 1994]. In this paper we study to which extent the independence assumption can be relaxed by replacing it with restrictions on higher order moment or cumulant tensors of ε. We document new conditions that establish identification for several non-independent component models, e.g. common variance models, and propose efficient estimation methods based on the identification results. We show that in situations where independence cannot be assumed the efficiency gains can be significant relative to methods that rely on independence.

1. INTRODUCTION

Consider the linear system

$$AY = \varepsilon \ , \tag{1}$$

where $Y \in \mathbb{R}^d$ is observed, $A \in \mathbb{R}^{d\times d}$ is invertible, and ε is a mean-zero hidden random vector with uncorrelated components. If ε is standard Gaussian, or more generally spherical, then the distribution of Y can identify A only up to orthogonal transformations. In contrast, if the components of ε are mutually independent and at least $d-1$ are non-Gaussian, then A can be identified up to permutation and sign transformations of its rows [Comon, 1994]. This result follows from the Darmois-Skitovich theorem [Darmois, 1953, Skitovic, 1953] and forms the building block of the vast literature on independent components analysis (ICA) [e.g. Hyvärinen et al., 2001b, Comon and Jutten, 2010, Hyvärinen, 2013].

As implied by its name, the working assumption in the ICA literature is that the components of ε are independent. For some applications this is an important starting principle as the interest is explicitly in recovering the independent components, see for instance the cocktail party problem described in Hyvärinen et al. [2001b, p. 148]. However, in other applications, where the interest is solely in recovering A, the independence assumption is not a crucial starting point and can in fact be restrictive as the distribution of Y may not admit a linear transformation that leads to independent components [e.g. Hyvärinen et al., 2001a, Matteson and Tsay, 2017].

Date: September 9, 2023.
2020 *Mathematics Subject Classifications.* 15A69, 62H99.

To this extent, in this paper we study assumptions that (i) relax the independence assumption yet (ii) assure the identifiability of the matrix A from observations of Y. We generally normalize $\text{var}(\varepsilon) = I_d$ which implies that $\text{var}(Y) = (A'A)^{-1}$ and narrows down the identification problem to the compact set $\Omega = \{QA : Q \in \text{O}(d)\}$, where $\text{O}(d)$ is the set of d-dimensional orthogonal matrices. This refinement allows to formally state our research question: Which higher order restrictions on ε allow to identify a finite, possibly structured, subset of Ω?

We systematically study our question by considering different restrictions on the higher order moments or cumulants of ε. We focus on cases where a subset of entries of a given rth order moment/cumulant tensor are set to zero, for some $r > 2$. Although there are alternative types of restrictions that can be considered, zero restrictions are attractive as they can often be justified by generative models for ε such as the common variance, scale-elliptical and mean-independent component models introduced in Section 2. Additionally, zero restrictions often arise naturally from subject specific knowledge, see Bekaert et al. [2021] for examples from economics.

We provide two classes of higher order restrictions that (a) identify the set of signed permutation matrices and (b) strictly relax the identification assumptions of Comon [1994].

First, we consider the class where the off-diagonal elements of a given moment or cumulant tensor are all zero. Such off-diagonal restrictions are often adopted for estimation in the ICA literature under the independence assumption.[1] We show that, without imposing the independence assumption, if we set the off-diagonal elements of *any* rth order moment or cumulant tensor to zero we obtain sufficient identifying restrictions to pin down Q up to sign and permutation. We point out that for $r = 3, 4$ similar results are shown for moment restrictions in Guay [2021] and Velasco [2022] using a different proof strategy, which does not generalize to higher r.

Second, while off-diagonal zero restrictions are commonly adopted, they cannot always be used when the components of ε are not independent. For instance, if ε follows a symmetric distribution the odd order tensors are all zero and provide no restrictions, but the even order tensors may not be diagonal as is the case, for instance, when the errors have common stochastic variance [e.g. Hyvärinen et al., 2001a, Montiel Olea et al., 2022]. This motivates our second class of tensor restrictions, which we refer to as reflectionally invariant restrictions, where the only non-zero tensor entries are those where each index appears even number of times. This provides a strict relaxation of the diagonal tensor assumption and we show that this assumption remains sufficient to identify Q up to sign and permutation.

[1] For instance the JADE algorithm of Cardoso and Souloumiac [1993] is based on diagonalizing the fourth order cumulant tensor.

Overall, diagonal and reflectionally invariant restrictions are most relevant for practical purposes, as efficient estimation methods can be easily implemented based on such identifying assumptions. Moreover, these restrictions allow to identify several specific non-independent components models, such as those with common variance components, scale-elliptical errors, and mean independent errors, for which previously no identification results existed.

With identification established we turn to estimation. For moment restrictions we note that generalized moment estimators [Hansen, 1982] are attractive as they are (i) easy to implement and (ii) semi-parametrically efficient in settings where the only known features of the model are the moment restrictions [e.g. Chamberlain, 1987]. We extend this class by also allowing for cumulant restrictions. The resulting class of higher order based minimum distance estimators is large and includes existing tensor based estimators for model (1), such as JADE [Cardoso and Souloumiac, 1993], as special cases, but also introduces new estimators. We show that estimators in this class are consistent and asymptotically normal under standard regularity conditions.

Our starting observation — independent components may not exists — is not new. In fact, such concerns were common in the early literature on Blind Source Separation, see Comon and Jutten [2010, Chapter 1] for an illuminating discussion, and they motivated explicit tests for the existence of independent components [e.g. Matteson and Tsay, 2017, Davis and Ng, 2022]. In addition, the possible absence of independent components motivated the usage of alternative identifying restrictions. For instance, a large literature has explored the usage of time/frequency characteristics of non-stationary components for identification [e.g. Comon and Jutten, 2010, Chapter 11]. In the current paper we do not exploit non-stationarity for identification.

There exists numerous methods for estimation and inference in independent components models: e.g. cumulant and moment based methods [Cardoso, 1989, Cardoso and Souloumiac, 1993, Cardoso, 1999, Hyvärinen, 1999, Lanne and Luoto, 2021, Drautzburg and Wright, 2021], kernel methods Bach and Jordan [2002], maximum likelihood methods Chen and Bickel [2006], Samworth and Yuan [2012], Lee and Mesters [2021] and rank based methods Ilmonen and Paindaveine [2011], Hallin and Mehta [2015]. Based on our new identification results these methods could be modified to relax the independence assumption. We perform this task for moment and cumulant based estimation methods, but clearly other methods could be modified as well. For moment estimators a well developed general inference theory exists, see Hall [2005] for a textbook treatment. For cumulant based estimators less work has been done. A notable exception is found for measurement error models where cumulant based estimators have been developed in Geary [1941] and Erickson et al. [2014]. The difference in their setting is that the parameters of interest can be written as a linear function of the higher order cumulants of the observables. For model (1) this is not possible.

The remainder of this paper is organized as follows. Section 2 provides motivating examples where independent components do not exist. Section 3 defines some tensor notation and reviews relevant existing results. The general problem that we study is introduced in Section 4. The new identification results are discussed in Section 5. Inference is discussed in Section 6 followed by some numerical results in Section 7. Any references to sections, equations, lemmas etc. which start with "S" refer to the supplementary material.

2. Examples of non-independent component models

Independent component analysis assumes that the components of the latent vector ε are completely independent. In this section we introduce a few examples of popular generative models for which the independence assumption is violated. Below we revisit these examples to show that these models do satisfy weaker higher order tensor restrictions based on which we can establish the identification of A up to permutation and sign.

2.1. **Common variance components models.** Consider

$$AY = \varepsilon, \qquad \text{with} \qquad \varepsilon = \tau\eta, \tag{2}$$

where τ is some positive random variable wth finite second moment and η is a random vector that is independent of τ and such that $\mathbb{E}(\eta) = 0$ and $\text{var}(\eta) = I_d$. In this situation

$$\text{var}(\varepsilon) = \text{var}(\mathbb{E}(\varepsilon|\tau)) + \mathbb{E}(\text{var}(\varepsilon|\tau)) = \mathbb{E}(\tau^2)I_d$$

and so the entries of ε remain uncorrelated. However, even if the components of η are independent, the components of ε are generally not. Indeed, assuming η has independent components, we have

$$\mathbb{E}(\varepsilon_i^2\varepsilon_j^2) = \mathbb{E}(\tau^4)\mathbb{E}(\eta_i^2)\mathbb{E}(\eta_j^2) = \mathbb{E}(\tau^4)$$

and

$$\mathbb{E}(\varepsilon_i^2)\mathbb{E}(\varepsilon_j^2) = \mathbb{E}(\tau^2)^2\mathbb{E}(\eta_i^2)\mathbb{E}(\eta_j^2).$$

Thus $\mathbb{E}(\varepsilon_i^2\varepsilon_j^2) \neq \mathbb{E}(\varepsilon_i^2)\mathbb{E}(\varepsilon_j^2)$, unless $\text{var}(\tau) = 0$, and A cannot be identified using the standard ICA assumptions.

In the ICA literature common variance models are one of the motivating examples for topographic ICA (TICA) Hyvärinen et al. [2001a], which can be used in image analysis Meyer-Base et al. [2003], Meyer-Bäse et al. [2004], among others. Further, in finance the variances of stock returns and other financial assets often depend on common components, see Asai et al. [2006] for a review of the literature. And while ICA has been applied in this context [e.g. Back and Weigend, 1997] the presence of common volatility limits its credibility. Finally, in macroeconomics there is also strong empirical evidence for common volatility structures [e.g. Ludvigson et al., 2021].

The non-independence for the baseline common variance model (2) carries over to more general models, where τ becomes a random vector. For

instance, let $K \in \mathbb{R}^{d\times m}$ be a fixed loading matrix, then a more general common variance model reads

$$AY = \varepsilon\,, \quad \text{with} \quad \varepsilon = \tau \odot \eta \quad \text{and} \quad \tau = \phi(KZ)\,, \tag{3}$$

where $\eta \in \mathbb{R}^d$ and $Z \in \mathbb{R}^m$ are independent random vectors with independent components, the function $\phi : \mathbb{R} \to \mathbb{R}_{>0} = \{x \in \mathbb{R} : x > 0\}$, is applied coordinatewise and $\odot$ denotes the Hadamard product. In this model $\tau \in \mathbb{R}^m$ is independent of η, and the components of ε share a common variance if they load on the same underlying components, or factors, Z. By exactly the same argument as above ε has uncorrelated components and, generally, $\mathbb{E}(\varepsilon_i^2\varepsilon_j^2) \neq \mathbb{E}(\varepsilon_i^2)\mathbb{E}(\varepsilon_j^2)$ and the ICA identification result does not apply.

Formal theoretical identifiability results for A along the lines of Comon [1994] and Eriksson and Koivunen [2003] have not been developed for common variance models. Section 5 provides such results.

2.2. **Scale elliptical components models.** Suppose that in the common variance model (2) the error $\varepsilon = \tau\eta$ satisfies $\eta = U \sim \mathcal{U}_d$, where $\mathcal{U}_d$ is the uniform distribution on the d-sphere. In this case the components of U are no longer independent and ε is said to follow an elliptical distribution [Kelker, 1970]. It follows that, generally, ε will not have independent components. The exception is the case where τ^2 follows a χ^2_d distribution, such that ε is standard normal.

In general, for elliptical errors A can never be recovered beyond the set $\Omega = \{QA : Q \in \mathrm{O}(d)\}$. This follows directly because the distribution of a spherical random vector is invariant under the orthogonal transformations. This limitation has been pointed out already e.g. in Palmer et al. [2007] who modify the Gaussian distribution to a distribution that is not rotationally invariant.

Here we follow the general idea of Forbes and Wraith [2014] and generalize the elliptical distribution by defining $\varepsilon = \tau\odot U$ with τ a d-dimensional vector. Such *multiple scale elliptical* distribution continues to have non-independent components but any variation in the components of τ will allow us to identify A using the higher order moment/cumulant restrictions introduced below. Moreover, this distribution has the attractive property that it allows to model different tail behavior in different dimensions; e.g. Gaussian in one dimension and Cauchy in another (see the discussion in Azzalini and Genton [2008]).

Formally, the multiple scale elliptical components model is given

$$AY = \varepsilon\,, \quad \text{with} \quad \varepsilon = \tau \odot U \quad \text{and} \quad U \sim \mathcal{U}_d\,, \tag{4}$$

with $\tau \in \mathbb{R}^d$ and U independent. The term *elliptical components analysis* was coined in Han and Liu [2018], but their interest was in recovering the top eigenvectors of AA' with $A = (a_1, \ldots, a_d)'$. Also, in the same model Vogel and Fried [2011] and Rossell and Zwiernik [2021] studied recovering AA'. These works were motivated by the robustness of the elliptical distribution

and its ability to preserve much of the dependence relationships offered by the Gaussian distribution.

In contrast, we are interested in recovering the full matrix A while taking advantage of the robustness of the elliptical distribution and the multiple scale generalization. However, as mentioned, in this case ε does not have independent components and we cannot rely on the classical ICA identification result. To circumvent this, Section 5 establishes new identification results for A in the multiple scale elliptical components model (4).

2.3. **Mean independent component models.** Besides specific generative models or probabilistic models, we can also define non-independent components models more directly by relaxing the strict independence requirement. Consider the following definition of a mean independent component model

$$a_i' Y \;=\; \varepsilon_i \;, \quad \text{with} \quad \mathbb{E}(\varepsilon_i | \varepsilon_{-i}) = 0 \;, \quad \text{for} \;\; i = 1, \ldots, d \;, \tag{5}$$

where ε_{-i} drops ε_i from ε. This relaxed notion of independence is attractive in practice as it avoids restricting terms like $\mathbb{E}(\varepsilon_i^k | \varepsilon_{-i})$ for $k = 2, 3, \ldots$, and for $\mathbb{E}(\varepsilon_i | \varepsilon_{-i}) = 0$ there often exist subject specific knowledge that can be used to justify the restriction.

To give a concrete example, suppose that $Y = (q, p)'$ where q is the quantity demanded of a good and p its price. In a baseline econometric model ε_1 and ε_2 are then known as demand and supply shocks [e.g Hayashi, 2000, Chapter 3]. Economic theory generally suggests that demand and supply shocks should not be able to predict each other, i.e. $\mathbb{E}(\varepsilon_1 | \varepsilon_2) = 0$ and vice versa. At the same time, restrictions of the form $\mathbb{E}(\varepsilon_1^k | \varepsilon_2) = 0$, for $k = 2, 3, \ldots$ are typically not motivated by economic theory. Additional economic motivation for specific moment conditions in supply and demand models is given in Bekaert et al. [2021, 2022].

In Section 5 we provide new identification results for the mean independent component model (5). The supplementary material Section S1 provides additional motivation for non-independent components models.

3. Basic tensor notation

Consider the random vector $X = (X_1, \ldots, X_d)'$ and let $M_X(\boldsymbol{t}) = \mathbb{E} e^{\boldsymbol{t}' X}$ and $K_X(\boldsymbol{t}) = \log \mathbb{E} e^{\boldsymbol{t}' X}$ denote the corresponding moment and cumulant generating functions, respectively. We write $\mu_r(X)$ to denote the r-order $d \times \cdots \times d$ moment tensor, that is an r-dimensional table whose $(i_1, \ldots, i_r)$-th entry is

$$\mu_r(X)_{i_1 \cdots i_r} \;=\; \mathbb{E} X_{i_1} \cdots X_{i_r} \;=\; \frac{\partial^r}{\partial t_{i_1} \cdots \partial t_{i_r}} M_X(\boldsymbol{t}) \Big|_{\boldsymbol{t}=0} \;.$$

Similarly, the cumulant tensor $\kappa_r(X)$ is defined as

$$\kappa_r(X)_{i_1 \cdots i_r} \;=\; \operatorname{cum}(X_{i_1}, \ldots, X_{i_r}) \;=\; \frac{\partial^r}{\partial t_{i_1} \cdots \partial t_{i_r}} K_X(\boldsymbol{t}) \Big|_{\boldsymbol{t}=0} \;.$$

We have $\kappa_1(X) = \mu_1(X)$, $\kappa_2(X) = \mu_2(X) - \mu_1(X)\mu_1'(X)$ and $\kappa_3(X)$ is a $d \times d \times d$ tensor filled with the third order central moments of X. The relationship between $\mu_r(X)$ and $\kappa_r(X)$ for higher order r is more cumbersome but very well understood Speed [1983], McCullagh [2018]; see the supplementary material S3.1. Directly by construction, $\mu_r(X)$ and $\kappa_r(X)$ are symmetric tensors, i.e. they are invariant under an arbitrary permutation of the indices. The space of real symmetric $d \times \cdots \times d$ order r tensors is denoted by $S^r(\mathbb{R}^d)$. Writing $[d] = \{1, \ldots, d\}$, the set of indices of an order r tensor is $[d]^r$. However, $S^r(\mathbb{R}^d) \subset \mathbb{R}^{d \times \cdots \times d}$ has dimension $\binom{d+r-1}{r}$ and the unique entries of $T \in S^r(\mathbb{R}^d)$ are $T_{i_1 \cdots i_r}$ for $1 \leq i_1 \leq \ldots \leq i_r \leq d$.

The vast majority of results in this paper holds for both moment and cumulant tensors. To avoid excessive notation we denote a given rth order moment or cumulant tensor by $h_r(X)$. Whenever distinguishing between moments or cumulants is required we specify towards $\mu_r(X)$ or $\kappa_r(X)$.

A critical feature of moment and cumulant tensors that we use to study identification in model (1) comes from multilinearity, i.e. for every $A \in \mathbb{R}^{d \times d}$ we have

$$h_r(AX) = A \bullet h_r(X), \tag{6}$$

where $A \bullet T$ for $T \in S^r(\mathbb{R}^d)$ denotes the standard multilinear action

$$(A \bullet T)_{i_1 \cdots i_r} = \sum_{j_1=1}^{d} \cdots \sum_{j_r=1}^{d} A_{i_1 j_1} \cdots A_{i_r j_r} T_{j_1 \cdots j_r}$$

for all $(i_1, \ldots, i_r) \in [d]^r$, see, for example, Section 2.3 in Zwiernik [2016].

Since $A \bullet T \in S^r(\mathbb{R}^d)$ for all $T \in S^r(\mathbb{R}^d)$ we say that $A \in \mathbb{R}^{d \times d}$ *acts* on $S^r(\mathbb{R}^d)$. The notation $A \bullet T$ is a special case of a general notation for multilinear transformations $\mathbb{R}^{n_1 \times \cdots \times n_r} \to \mathbb{R}^{m_1 \times \cdots \times m_r}$ given by matrices $A^{(1)} \in \mathbb{R}^{m_1 \times n_1}$, $\ldots$, $A^{(r)} \in \mathbb{R}^{m_r \times n_r}$:

$$[(A^{(1)}, \ldots, A^{(r)}) \cdot T]_{i_1 \cdots i_r} = \sum_{j_1=1}^{n_1} \cdots \sum_{j_r=1}^{n_r} A^{(1)}_{i_1 j_1} \cdots A^{(r)}_{i_r j_r} T_{j_1 \cdots j_r}. \tag{7}$$

See, for example Lim [2021] for an overview of the computational aspects of tensors.

Remark 3.1. The multilinearity property (6) is not exclusive to moments and cumulant tensors, as central moments, free cumulants and boolean cumulants, for instance, also share this property; see Zwiernik [2012, Section 5.2] for a more complete characterization. Our main results rely only on the property (6) and so the definition of $h_r(X)$ can be extended beyond moments and cumulants if needed.

The following well-known characterization of independence is of importance in our work.

Proposition 3.2. *The components of X are independent if and only if $\kappa_r(X)$ is a diagonal tensor for every $r \geq 2$.*

This result highlights that the necessity of the independence assumption in ICA can be investigated by studying the consequences of making appropriate higher order cumulant tensors elements non-zero. The relationship to the Gaussian distribution can be understood from a version of the Marcinkiewicz classical result Marcinkiewicz [1939], Lukacs [1958].

Proposition 3.3. *If* $X \sim \mathcal{N}_d(\mu, \Sigma)$ *then* $\kappa_1(X) = \mu$, $\kappa_2(X) = \Sigma$, *and* $\kappa_r(X) = \mathbf{0}$ *for* $r \geq 3$. *Moreover, the Gaussian distribution is the only probability distribution such that there exists* r_0 *with the property that* $\kappa_r(X) = \mathbf{0}$ *for all* $r \geq r_0$.

As we formalize below, this result implies that we require deviations from the Gaussian distribution to ensure identification in model (1), similar as required in the classical ICA result [Comon, 1994].

4. Identification with zero constraints

Since $AY = \varepsilon$ with $\mathbb{E}\varepsilon = 0$ and $\mathrm{var}(\varepsilon) = I_d$, the variance of Y satisfies $\mathrm{var}(Y) = (A'A)^{-1}$ and so it is enough to narrow down potential candidates for A to the compact set

$$\Omega \;:=\; \{QA : Q \in \mathrm{O}(d)\} \;.$$

Our main insight is as follows: Since ε is unobserved, multiplying (1) by $Q \in \mathrm{O}(d)$ gives an alternative representation $\widetilde{A}Y = \tilde{\varepsilon}$, where $\mathbb{E}\tilde{\varepsilon} = 0$ and $\mathrm{var}(\tilde{\varepsilon}) = I_d$. The goal is to define suitable additional restrictions on the distribution of ε so that the distribution of $\tilde{\varepsilon} = Q\varepsilon$ does not satisfy these restrictions unless Q is very special. The main result of [Comon, 1994] proposes to use non-Gaussianity and independence. We show how to exploit additional structure in some $h_r(X)$ to obtain similar results.

4.1. Exploiting general constraints.

Suppose that we have some additional information about a fixed higher-order tensor $T = h_r(\varepsilon) \in S^r(\mathbb{R}^d)$, for example we know that $T \in \mathcal{V}$ for some subset $\mathcal{V} \subseteq S^r(\mathbb{R}^d)$. By multilinearity (6) we have

$$T = h_r(AY) = A \bullet h_r(Y) \;, \tag{8}$$

and for any given $Q \in \mathrm{O}(d)$, $QA \in \Omega$ remains a valid candidate if

$$(QA) \bullet h_r(Y) \;\in\; \mathcal{V} \;. \tag{9}$$

However,

$$(QA) \bullet h_r(Y) \;=\; Q \bullet (A \bullet h_r(Y)) \;=\; Q \bullet T$$

and so (8) and (9) hold together if and only if $Q \bullet T \in \mathcal{V}$. For $T \in \mathcal{V}$, we define

$$\mathcal{G}_T(\mathcal{V}) \;:=\; \{Q \in \mathrm{O}(d) : Q \bullet T \in \mathcal{V}\} \;, \tag{10}$$

which is the subset of Ω that can be identified from $\mathcal{V}$. Below we sometimes drop $\mathcal{V}$, writing $\mathcal{G}_T$, if the context is clear. We always have $I_d \in \mathcal{G}_T(\mathcal{V})$ but in general $\mathcal{G}_T(\mathcal{V})$ will be larger.

We summarize the general identification problem as follows.

Proposition 4.1. *Consider the model (1) with* $\mathbb{E}\varepsilon = 0$ *and* $\mathrm{var}(\varepsilon) = I_d$. *Suppose we know, for a fixed* $r \geq 3$, *that* $T = h_r(\varepsilon) \in \mathcal{V} \subset S^r(\mathbb{R}^d)$. *Then* A *can be identified up to the set*

$$\Omega_0 \;=\; \{QA : Q \in \mathcal{G}_T(\mathcal{V})\}. \tag{11}$$

In the ideal situation $\mathcal{G}_T(\mathcal{V})$ is a singleton, in which case A can be recovered exactly. But we also expect that, in general, exact recovery will not be possible. We are therefore looking for restrictions $\mathcal{V}$ that assure that $\mathcal{G}_T(\mathcal{V})$ is a finite set, possibly with some additional structure. The leading structure of interest is the set of signed permutations for which we recover the original ICA result under strictly weaker assumptions. We denote the set of $d \times d$ signed permutation matrices by $\mathrm{SP}(d)$. These are the $2^d d!$ matrices that are of the form DP, where $D, P \in \mathrm{O}(d)$ with D diagonal and P a permutation matrix.

4.2. **Zero restrictions.** Clearly, there exists a plethora of restrictions on the higher order moment or cumulants that can be considered. For instance, the ICA assumption imposes that $\kappa_r(\varepsilon) = \mathrm{cum}_r(\varepsilon)$ has zero off-diagonal elements for all r (i.e. Proposition 3.2). At the same time we know from the discussion in Section 2 that several generative models do not satisfy these restrictions and hence we seek relaxations that accommodate such models yet still yield identification of A.

We formalize zero restrictions by choosing a subset $\mathcal{I}$ of r-tuples $(i_1, \ldots, i_r)$ satisfying $1 \leq i_1 \leq \cdots \leq i_r \leq d$ and by defining the vector space $\mathcal{V} = \mathcal{V}(\mathcal{I})$ of symmetric tensors $T \in S^r(\mathbb{R}^d)$ such that $T_{\boldsymbol{i}} = 0$ for all $\boldsymbol{i} = (i_1, \ldots, i_r) \in \mathcal{I}$. In symbols:

$$\mathcal{V} \;=\; \mathcal{V}(\mathcal{I}) \;=\; \{T \in S^r(\mathbb{R}^d) :\; T_{\boldsymbol{i}} = 0 \text{ for } \boldsymbol{i} \in \mathcal{I}\}\;.$$

Note that the codimension of $\mathcal{V}$ in $S^r(\mathbb{R}^d)$ is precisely $\mathrm{codim}(\mathcal{V}) = |\mathcal{I}|$.

The following example clarifies our notation and illustrates how higher order moment or cumulant restrictions can be used for identification.

Example 4.2. Suppose that $\mathcal{V} \subset S^3(\mathbb{R}^2)$ is given by $T_{112} = T_{122} = 0$. This is a two-dimensional subspace parametrized by T_{111} and T_{222}. The condition $Q \bullet T \in \mathcal{V}$ is given by the system of two cubic equations in the entries of Q

$$\begin{aligned} Q_{11}^2 Q_{21} T_{111} + Q_{12}^2 Q_{22} T_{222} \;&=\; 0 \\ Q_{11} Q_{21}^2 T_{111} + Q_{12} Q_{22}^2 T_{222} \;&=\; 0. \end{aligned}$$

In a matrix form this can be written as

$$Q \cdot \begin{bmatrix} Q_{11} & 0 \\ 0 & Q_{22} \end{bmatrix} \cdot \begin{bmatrix} Q_{21} & 0 \\ 0 & Q_{12} \end{bmatrix} \cdot \begin{bmatrix} T_{111} \\ T_{222} \end{bmatrix} \;=\; \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Since Q is orthogonal, each of the two diagonal matrices above is either identically zero or it is invertible. If it is identically zero then Q must be a sign permutation matrix and the equation clearly holds. If they are

both invertible we immediately see that the equation cannot hold unless $T_{111} = T_{222} = 0$, in which case T is the zero tensor showing that for every nonzero $T \in \mathcal{V}$ we have that $\mathcal{G}_T(\mathcal{V}) = \mathrm{SP}(2)$.

Unfortunately, the direct arguments that we used in this example to determine $\mathcal{G}_T(\mathcal{V})$ do not generalize for higher r and d. Handling such cases requires a more systematic approach which we develop in Section 5.

Remark 4.3. For exposition purposes we only consider cases where some entries of $T = h_r(\varepsilon)$ are set to zero, but we note that our results can be extended for cases where entries of T are non-zero but *known* to the researcher. An example with 4th order moment restrictions arises when $\mathbb{E}\varepsilon = 0$, $\mathrm{var}(\varepsilon) = I_d$ and $T_{iijj} = \mathbb{E}\varepsilon_i^2\varepsilon_j^2 = 1$ for $i \neq j$.

5. Identification up to Sign and Permutation

In this section we discuss specific sets of zero restrictions that allow to identify A up to sign and permutation. For each specific set of zero restrictions we give concrete examples of models that can be identified using such restrictions.

5.1. Diagonal tensors.

Denote by $T = h_r(\varepsilon)$ the rth order moment or cumulant tensor of ε. A simple assumption that facilitates identification is that T is a diagonal tensor.

Definition 5.1. A tensor $T \in S^r(\mathbb{R}^d)$ is called diagonal if it has entries $T_{\boldsymbol{i}} = 0$ unless $\boldsymbol{i} = (i, \ldots, i)$ for some $i = 1, \ldots, d$.

Of course, if the components of ε are independent then $\kappa_r(\varepsilon)$ is diagonal for all $r \geq 2$ (see Proposition 3.2). Assuming that T is diagonal is much less restrictive than full independence as any T can be chosen without imposing restrictions on other cumulants, or moments. This allows for instance to assume that only the cross-third moments of ε are zero, without imposing any restrictions on the higher order moments.

In this section, $\mathcal{V}$ denotes the set of diagonal tensors in $S^r(\mathbb{R}^d)$. For verifying whether $\mathcal{V}$ provides sufficient identifying restrictions we will study the tensors T and $Q \bullet T$ via their associated homogeneous polynomials in variables $x = (x_1, \ldots, x_d)$. We have

$$f_T(x) = \sum_{i_1=1}^{d} \cdots \sum_{i_r=1}^{d} T_{i_1 \ldots i_r} x_{i_1} \ldots x_{i_r} \;=\; \sum_{\boldsymbol{i}} T_{\boldsymbol{i}} (x^{\otimes r})_{\boldsymbol{i}} \;=\; \langle T, x^{\otimes r} \rangle, \tag{12}$$

where $x^{\otimes r} \in S^r(\mathbb{R}^d)$ denotes the tensor with coordinates $(x^{\otimes r})_{i_1 \cdots i_r} = x_{i_1} \cdots x_{i_r}$. If $r = 2$ then T is a symmetric matrix and $f_T(x) = x'Tx$ is the standard quadratic form associated with T.

Lemma 5.2. *If $T \in S^r(\mathbb{R}^d)$ and $A \in \mathbb{R}^{d \times d}$ then $f_{A \bullet T}(x) = f_T(A'x)$. Moreover, $\nabla f_{A \bullet T} = A \nabla f_T(A'x)$ and $\nabla^2 f_{A \bullet T} = A \nabla^2 f_T(A'x) A'$.*

Proof. The first claim follows because

$$f_{A\bullet T}(x) \;=\; \langle A \bullet T, x^{\otimes r}\rangle \;=\; \langle T, (A'x)^{\otimes r}\rangle \;=\; f_T(A'x).$$

The second claim is then a direct check. □

This will be useful for deriving our first main result.

Theorem 5.3. *Let $T \in S^r(\mathbb{R}^d)$ for $r \geq 3$ be a diagonal tensor with at most one zero entry on the diagonal. Then $Q \bullet T \in \mathcal{V}$ if and only if $Q \in \mathrm{SP}(d)$, i.e. $\mathcal{G}_T(\mathcal{V}) = SP(d)$.*

Proof. The left direction is clear. For the right direction, note that the tensor T is diagonal if and only if $\nabla^2 f_T(x)$ is a diagonal polynomial matrix. By Lemma 5.2, we have $f_{Q\bullet T}(x) = f_T(Q'x)$ and

$$\nabla^2 f_{Q\bullet T}(x) \;=\; Q\nabla^2 f_T(Q'x)Q'.$$

Thus, $Q \bullet T$ is diagonal if and only if $Q\nabla^2 f_T(Q'x)Q' = D(x)$ for a diagonal matrix $D(x)$. Equivalently, for every i, j

$$Q_{ij}\frac{\partial^2}{\partial x_j^2} f_T(Q'x) = D_{ii}(x)Q_{ij},$$

where we also used the fact that $\nabla^2 f_T(x)$ is a diagonal matrix. If each row of Q has exactly one non-zero entry then $Q \in \mathrm{SP}(d)$ and we are done. So suppose $Q_{ij}, Q_{ik} \neq 0$. Then, by the above equation

$$\frac{\partial^2}{\partial x_j^2} f_T(Q'x) \;=\; D_{ii}(x) \;=\; \frac{\partial^2}{\partial x_k^2} f_T(Q'x).$$

This is an equality of polynomial functions and thus, equivalently, $\frac{\partial^2}{\partial x_j^2} f_T(x) = \frac{\partial^2}{\partial x_k^2} f_T(x)$, which simply states that

$$T_{j\cdots j} x_j^{r-2} \;=\; T_{k\cdots k} x_k^{r-2}.$$

Since $r \geq 3$, this equality can hold only if $T_{j\cdots j} = T_{k\cdots k} = 0$, which is impossible by our genericity assumption. □

Remark 5.4. The genericity condition is a necessary condition. Indeed, if, for example $T_{1\cdots 1} = T_{2\cdots 2} = 0$ then $\frac{\partial}{\partial x_1^2} f_T(x) = \frac{\partial^2}{\partial x_2^2} f_T(x) = 0$. Thus, $\nabla^2 f_{Q\bullet T}(x)$ is diagonal for any block matrix of the form

$$Q \;=\; \begin{bmatrix} Q_0 & 0 \\ 0 & I_{d-2} \end{bmatrix}$$

where $Q_0 \in O(2)$ is an orthogonal matrix. The family of such matrices is infinite.

Combining Proposition 4.1 and Theorem 5.3 implies the following result.

Theorem 5.5. *Consider the model (1) with $\mathbb{E}\varepsilon = 0$, $\mathrm{var}(\varepsilon) = I_d$ and suppose that for some $r \geq 3$ the tensor $h_r(\varepsilon)$ is diagonal with at most one zero on the diagonal. Then A in (1) is identifiable up to permuting and swapping signs of its rows.*

Remark 5.6. Note that if ε is standard Gaussian then all higher order cumulants vanish. All odd-order moments vanish too. Even-order moment tensors are not zero but they are also not diagonal.

Subsequently, based on Theorem 5.5 we can provide an identification result for the common variance and mean independent component models that were discussed in Section 2.

Corollary 5.7. *Consider the model $AY = \varepsilon$ with $\mathbb{E}\varepsilon = 0$ and $\mathrm{var}(\varepsilon) = I_d$. Suppose that additionally ε satisfies one of the following conditions.*

(a) *$\varepsilon = \tau \odot \eta$, $\tau = \phi(KZ)$, where $K \in \mathbb{R}^{d\times m}$ is a fixed matrix, $\eta \in \mathbb{R}^d$ and $Z \in \mathbb{R}^m$ are independent random vectors with independent components and the function $\phi : \mathbb{R} \to \mathbb{R}_{>0}$, is applied coordinatewise.*

(b) *$\mathbb{E}(\varepsilon_i|\varepsilon_{-i}) = 0$ for $i = 1, \ldots, d$.*

Then $h_3(\varepsilon) = \mu_3(\varepsilon) = \kappa_3(\varepsilon)$ is a diagonal tensor. If additionally, $h_r(\varepsilon)$ has at most one zero on the diagonal then A is identifiable up to permuting and swapping signs of its rows.

Proof. The proof is based on Theorem 5.5. For (a) note that $h_3(\varepsilon)$ is diagonal as $\mathbb{E}\varepsilon_i\varepsilon_j\varepsilon_k = \mathbb{E}\tau_i\tau_j\tau_k\mathbb{E}\eta_i\eta_j\eta_k = 0$ unless $i = j = k$. For (b), consider the triple (i, j, k). Unless $i = j = k$, there will be at least one element that appears only once. Without loss of generality assume $i \neq j$ and $i \neq k$. We have

$$\mathbb{E}\varepsilon_i\varepsilon_j\varepsilon_k = \mathbb{E}(\mathbb{E}(\varepsilon_i\varepsilon_j\varepsilon_k|\varepsilon_{-i})) = \mathbb{E}(\varepsilon_j\varepsilon_k\mathbb{E}(\varepsilon_i|\varepsilon_{-i})) = 0$$

again confirming that the third order moment/cumulant tensor is diagonal. □

Our result for diagonal tensors cannot be used for the scaled elliptical distributions in (4). In this case all odd-order moments/cumulants are zero (not generic) and the even-order moment/cumulant tensors are not diagonal. This motivates our next section.

5.2. **Reflectionally invariant tensors.** In some applications the assumption that T is diagonal may be unattractive. A leading example is the scale elliptical components models where the third order tensors are zero but the fourth order tensor is not diagonal as entries of the form T_{iijj} cannot be restricted to zero (or some other constant). Note that the latter zero restriction is also invalid in the common variance model (3).

These observations motivate the following tensor restrictions.

Definition 5.8. A tensor $T \in S^r(\mathbb{R}^d)$ is called *reflectionally invariant* if the only potentially non-zero entries in T are the entries $T_{i_1\cdots i_r}$ where each

index appears in the sequence $(i_1, \dots, i_r)$ even number of times. If r is odd, the only reflectionally invariant tensor is the zero tensor.

To prove that reflectionally invariant tensors can be used to identify A in (1), recall from (12) that any $T \in S^r(\mathbb{R}^d)$ has an associated homogeneous polynomial $f_T(x)$ of order r in $x = (x_1, \dots, x_d)$. It is clear from the definition that a non-zero $T \in S^r(\mathbb{R}^d)$ is reflectionally invariant if and only if r is even and there is a homogeneous polynomial g_T of order $l := r/2$ such that $f_T(x) = g_T(x_1^2, \dots, x_d^2)$. We have the following useful characterization of reflectionally invariant tensors.

Lemma 5.9. *The tensor $T \in S^r(\mathbb{R}^d)$ is reflectionally invariant if and only if $f_T(x) = f_T(Dx)$ for every diagonal matrix with ± 1 on the diagonal.*

Proof. By Lemma 5.2, $f_T(x) = f_T(Dx)$ is equivalent to saying that $D \bullet T = T$ for every diagonal $D \in \mathbb{Z}_2^d$. If T is reflectionally invariant then

$$f_T(Dx) = g_T(D_{11}^2 x_1^2, \dots, D_{dd}^2 x_d^2) = f_T(x),$$

which establishes the right implication. For the left implication note that $f_T(x) = f_T(Dx)$, for each D, implies that f_T does not depend on the signs of the components of x. Since this is a polynomial, we must be able to write it in the form $g_T(x_1^2, \dots, x_d^2)$ (this is obvious in one dimension and, in general, can be proved in each dimension separately). This is equivalent with T being reflectionally invariant. □

In the theorem below, for a tensor $T \in S^r(\mathbb{R}^d)$ we use the notation

$$T_{+\cdots+ij} := \sum_{i_1=1}^{d} \cdots \sum_{i_{r-2}=1}^{d} T_{i_1 \cdots i_{r-2} ij}.$$

Theorem 5.10. *Suppose that $T \in S^r(\mathbb{R}^d)$ for an even r is a reflectionally invariant tensor satisfying*

$$T_{+\cdots+ii} \neq T_{+\cdots+jj} \quad \text{for all } i \neq j. \tag{13}$$

Then $Q \bullet T$ is reflectionally invariant for $Q \in \mathrm{O}(d)$ if and only if $Q \in \mathrm{SP}(d)$, i.e. $\mathcal{G}_T(\mathcal{V}) = SP(d)$.

Remark 5.11. We emphasize that the genericity condition in (13) simply states that T lies outside of $\binom{d}{2}$ explicit linear hyperplanes in $S^r(\mathbb{R}^d)$. It is interesting to observe how the genericity condition evolves when zero restrictions are relaxed. First, in the classical ICA result [Comon, 1994] the condition is that for each $i = 1, \dots, d$ the corresponding diagonal entries of $h_r(\varepsilon)$ across r cannot all be zero. The diagonal tensor identification result (Theorem 5.5) replaces this condition by the requirement that at most one diagonal entry for *a given* $h_r(\varepsilon)$ can be zero. Finally, the reflectional invariant condition (13) extends the condition to the specific $\binom{d}{2}$ hyperplanes of $S^r(\mathbb{R}^d)$ which include the previous genericity conditions as isolated points.

Theorem 5.10 is proven using the following lemma.

Lemma 5.12. *Let r be even and suppose that $T \in S^r(\mathbb{R}^d)$ is reflectionally invariant tensor satisfying (13). Then $Q \bullet T = T$ for $Q \in \mathrm{O}(d)$ if and only if Q is a diagonal matrix.*

Proof. The left implication is clear because $f_T(x) = f_T(Dx) = f_{D\bullet T}(x)$ by Lemma 5.9. We prove the right implication by induction. The base case is $r = 2$, where the set of reflectionally invariant tensors corresponds to diagonal matrices. In this case the equation $Q \bullet T = T$ becomes $QTQ' = T$ or, equivalently, $QT = TQ$. This implies that for each $1 \le i \le j \le d$

$$Q_{ij}T_{jj} = T_{ii}Q_{ij}.$$

By the genericity condition (13), all the diagonal entries of the matrix T are distinct. In this case, for every $i \neq j$, we necessarily have $Q_{ij} = 0$. Proving that Q must be diagonal. Note also that this genericity condition is necessary: If two diagonal entries of T are equal, then the entries of the 2×2 submatrix $Q_{ij,ij}$ are not constrained, so Q does not have to be diagonal.

Suppose now that the claim is true for $r \ge 2$ and let $T \in S^{r+2}(\mathbb{R}^d)$ with $Q \bullet T = T$. Rewrite $Q \bullet T = T$, using the general multilinear notation (7), as

$$(Q, \dots, Q, I_d, I_d) \cdot T \;=\; (I_d, \dots, I_d, Q', Q') \cdot T. \tag{14}$$

We want to show that this equality implies that Q is a diagonal matrix. Let $\boldsymbol{i} = (i_1, \dots, i_r)$ and consider all $(r+2)$-tuples $(\boldsymbol{i}, u, u)$ for some $u \in \{1, \dots, d\}$. Writing (14) restricted to these indices gives

$$\sum_{j_1,\dots,j_r} Q_{i_1j_j} \cdots Q_{i_rj_r} T_{j_1\cdots j_r uu} \;=\; \sum_{j_{r+1},j_{r+2}} Q_{j_{r+1}u} Q_{j_{r+2}u} T_{i_1\cdots i_r j_{r+1} j_{r+2}}.$$

Now sum both sides over all $u = 1, \dots, d$. Using the fact that Q is orthogonal we get that $\sum_u Q_{j_{r+1}u}Q_{j_{r+2}u}$ is zero if $j_{r+1} \neq j_{r+2}$ and it is 1 if $j_{r+1} = j_{r+2}$. Denoting $S_{\boldsymbol{i}} = \sum_u T_{\boldsymbol{i}uu}$, summation over u yields

$$\sum_{j_1,\dots,j_r} Q_{i_1j_j} \cdots Q_{i_rj_r} S_{j_1\cdots j_r} \;=\; \sum_{v} T_{i_1\cdots i_r vv} \;=\; S_{i_1\cdots i_r}.$$

Since this equation holds for every $\boldsymbol{i} = (i_1, \dots, i_r)$, we conclude $Q \bullet S = S$, where $S = (S_{\boldsymbol{i}}) \in S^r(\mathbb{R}^d)$. Note however that S is a reflectionally invariant tensor. Indeed, if some index appears in $\boldsymbol{i}$ odd number of times then $S_{\boldsymbol{i}} = T_{\boldsymbol{i}uu} = 0$ as the same index appears in $(\boldsymbol{i}, u, u)$ odd number of times. Since T satisfies (13), S satisfies (13) too. Indeed,

$$\sum_{k_1} \cdots \sum_{k_{l-1}} S_{k_1k_1\cdots k_{l-1}k_{l-1}ii} = \sum_{k_1} \cdots \sum_{k_{l-1}} \sum_{k_l} T_{k_1k_1\cdots k_{l-1}k_{l-1}k_lk_lii}$$

and so these quantities are distinct for all $i = 1, \dots, d$ by assumption on T. Now, by the induction assumption, we conclude that Q is diagonal. □

Proof of Theorem 5.10. The left implication is clear. For the right implication, suppose $Q \in \mathrm{O}(d)$ is such that $Q \bullet T$ is reflectionally invariant. By

Lemma 5.9, equivalently, $f_{Q\bullet T}(x) = f_{Q\bullet T}(Dx)$ for every diagonal $D \in \mathrm{O}(d)$, which gives $f_T(Q'x) = f_T(Q'Dx)$. This polynomial equation implies that

$$f_T(x) = f_T(Q'DQx)$$

but since $Q'DQ \in O(d)$, Lemma 5.12 implies that $\bar{D} = Q'DQ$ must be diagonal. Therefore, the equation $DQ = Q\bar{D}$ shows that switching the signs in the i-th row of Q is equivalent to switching some columns of Q. Suppose that there are at least two non-zero entries Q_{ik}, Q_{il} in the i-th row of Q and let D be such that $D_{ii} = -1$ and $D_{jj} = 1$ for $j \neq i$. The equality $DQ = Q\bar{D}$ requires that $\bar{D}_{kk} = \bar{D}_{ll} = -1$ and that Q has no other non-zero entries in k-th and l-th columns. Since these columns are orthogonal we get a contradiction concluding that the i-th row of Q must contain at most (and so exactly) one non-zero entry. Applying this to each $i = 1, \ldots, d$, we conclude that $Q \in \mathrm{SP}(d)$. □

Combining Proposition 4.1 and Theorem 5.10 implies the following result.

Theorem 5.13. *Consider the model (1) with $\mathbb{E}\varepsilon = 0$, $\mathrm{var}(\varepsilon) = I_d$ and suppose that for some even r the tensor $h_r(\varepsilon)$ is reflectionally invariant and it satisfies the genericity condition (13). Then A is identifiable up to permuting and swapping signs of its rows.*

Using Theorem 5.13 we can provide identification results for all the models from Section 2. We summarize these all in the following statement.

Corollary 5.14. *Consider the model $AY = \varepsilon$ with $\mathbb{E}(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) = I_d$. If ε follows the general common covariance model (3) then $h_4(\varepsilon)$ is reflectionally invariant. If ε follows the the multiple scaled elliptical distribution (4), then $h_r(\varepsilon)$ is reflectionally invariant for every even $r \geq 4$. If ε is mean independent as in (5) then $h_4(\varepsilon)$ is reflectionally invariant. If additionally, the genericity condition (13) holds then A is identifiable up to permuting and swapping signs of its rows.*

Proof. For the first statement consider any element of $h_4(\varepsilon)$ such that one index appears odd number of times. Then we can assume it appears exactly once. So let $j, k, l \neq i$ and then

$$\mathbb{E}(\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_l) = \mathbb{E}(\eta_i)\mathbb{E}(\eta_j\eta_k\tau_i\tau_j\tau_k) = 0.$$

Similar calculations hold for cumulants. For the second statement, observe that if D_i is the diagonal matrix with -1 on the (i,i)-th entry and 1 on the remaining diagonal entries, then

$$D_i(\tau \odot U) \;=\; \tau \odot (D_iU) \;\stackrel{d}{=}\; \tau \odot U, \tag{15}$$

where U, τ is like in (4). This assures invariance of the distribution (and so also the moments/cumulants) with respect to sign swapping of single coordinates. By Lemma 5.9, all moment/cumulant tensors must be reflectionally

invariant. For the third statement we start as for the first. So let $j, k, l \neq i$ and then

$$\mathbb{E}(\varepsilon_i \varepsilon_j \varepsilon_k \varepsilon_l) = \mathbb{E}(\varepsilon_j \varepsilon_k \varepsilon_l \mathbb{E}(\varepsilon_i | \varepsilon_{-i})) = 0$$

with similar calculations for cumulants. □

Note that if ε is standard normal (which arises as a degenerate case of both (3) and (4)) then the even-order moment tensors are non-zero reflectionally invariant tensors. However, they are not generic in the sense of (13). For example, if $r = 4$ then $\mathbb{E}\varepsilon_i^4 = 3$ and $\mathbb{E}\varepsilon_i^2 \varepsilon_j^2 = 1$ for $i \neq j$. This gives $T_{++ii} = d + 2$ for all $i = 1, \ldots, d$.

Remark 5.15. The genericity condition (13) can be worked out explicitly for each model. For instance, for the general common variance model (a) we have for $\mu_4(\varepsilon)$ that

$$\sum_{l=1}^{d} \mathbb{E}(\tau_l^2 \tau_i^2) \mathbb{E}(\eta_l^2 \eta_i^2) \neq \sum_{l=1}^{d} \mathbb{E}(\tau_l^2 \tau_i^2) \mathbb{E}(\eta_l^2 \eta_i^2) \qquad \text{for all } i \neq j \ .$$

This simplifies in the case for the simple common variance model (2) to become $\mathbb{E}(\eta_i^4) \neq \mathbb{E}(\eta_j^4)$ for all $i \neq j$. In the multiple scaled elliptical model we can exploit the summetry in the moments of $\eta \sim \mathcal{U}_d$ to show that $\mu_4(\varepsilon)$ is generic as long as

$$2\mathbb{E}\tau_i^4 + \sum_{k \neq i} \mathbb{E}(\tau_i^2 \tau_k^2) \ \neq \ 2\mathbb{E}\tau_j^4 + \sum_{k \neq j} \mathbb{E}(\tau_j^2 \tau_k^2) \qquad \text{for all } i \neq j.$$

5.3. **Generalizations.** Theorems 5.5 and 5.13 highlight key zero moment and cumulant patterns that can be used to identify A up to sign and permutation for the general model (1). Such restrictions are equally sufficient for identification in the class of linear simultaneous equations models $AY = BX + \varepsilon$ when X is exogenous, and various dynamic extensions of such models [e.g. Kilian and Lütkepohl, 2017].

That said it is also of interest to explore whether relaxing additional zero restrictions still leads to identification (up to sign and permutation), and which genericity conditions are required. Since $\dim(\mathrm{O}(d)) = \binom{d}{2}$, we need at least that many constraints to assure $\mathcal{G}_T$ is finite. However, as we formally show in the supplementary material Section S2 the minimal set

$$\mathcal{I} \ = \ \{(i, j, \ldots, j) : \ \ 1 \leq i < j \leq d\} \ ,$$

implies that $\mathcal{G}_T$ is finite, but in general $\mathcal{G}_T \neq \mathrm{SP}(d)$. Example S8 explicitly computes the difference between $\mathcal{G}_T$ and $\mathrm{SP}(d)$ for the illustrative case with $r = 3$ and $d = 2$. This finding has important implication that it is, in general, not sufficient to prove that the Jacobian of the moment or cumulant restrictions is full rank in order to establish that the identified set is equal to the set of signed permutations.

Motivated by these calculations, consider a special model with

$$\mathcal{I} \ = \ \{(i, j, \ldots, j) : 1 \leq i < j \leq d\} \cup \{(i, \ldots, i, j) : 1 \leq i < j \leq d\} \ .$$

Conjecture 5.16. If T is a generic tensor in $\mathcal{V}(\mathcal{I})$ then $Q \bullet T \in \mathcal{V}(\mathcal{I})$ if and only if $Q \in \mathrm{SP}(d)$.

The case when $d = 2$ is very special because $\mathrm{O}(2)$ has dimension 1. In this case the analysis of zero patterns can be often done using classical algebraic geometry tools. In particular, we can show that the conjecture holds for $S^r(\mathbb{R}^2)$ tensors for any $r \geq 3$.

Proposition 5.17. *Suppose that $T \in S^r(\mathbb{R}^2)$ satisfies $T_{12\cdots 2} = T_{1\cdots 12} = 0$ but is otherwise generic. Then $Q \bullet T \in \mathcal{V}(\mathcal{I})$ if and only if $Q \in \mathrm{SP}(2)$.*

We prove this result in Appendix S7.1. The genericity conditions are again linear and can be recovered from the proof.

Remark 5.18. Consider the two-dimensional supply and demand model described in Section 2.3. Proposition 5.17 assures that the matrix A can be identified up to scaling and swapping rows from $h_r(\varepsilon)$ for any $r \geq 3$ as long as it satisfies the genericity conditions.

5.3.1. *Towards point identification.* As stated, the results so far provided identification results for A up to the set of signed-permutation matrices. In practice, it is often of interest to reduce this set further to, perhaps, a singleton. Restricting additional higher order tensors to zero cannot help with this objective, as even the most stringent selection of zero restrictions, i.e. all higher order tensors are diagonal (the independent components case), only yields identification up to signed permutations.

Therefore, we briefly mention a few existing routes that different strands of literature have adopted for further shrinking the identified set. First, topographical ICA, which was motivated by the common variance model, suggests to explicitly model the common variance structure, i.e. model K in (3). It then imposes the additional assumption that only nearby latent components have the higher order dependency in order to pin down a unique permutation [e.g. Hyvärinen et al., 2001a]. Second and related, Shimizu et al. [2006] impose the additional assumption that there exists a permutation A that is lower triangular. This restriction further pins down the identified set up to sign changes. Moreover, the implied directed acyclic graphical model has a causal interpretation. Third, in econometrics sign restrictions on A are a popular tool to weed out economically uninteresting permutations of A. A canonical case arises when model (1) represents a demand and supply equation (cf Section 2.3), in which case it is natural to impose that the demand equation has a downward slope and the supply equation an upward slope. Together with the normalization that the scales on the errors are positive yields a unique permutation. Such schemes can be generalized for larger values of d.

6. Inference for non-independent components models

Given our new identification results, there exist numerous possible routes for estimating A in $AY = \varepsilon$ given a sample $\{Y_s\}_{s=1}^n$. A natural approach is

based on the concept of minimum distance estimation, where (i) the identifying zero restrictions from Section 5 are replaced by their sample equivalents and (ii) the parameter A is chosen such that the sample restrictions are as close as possible to zero, i.e. the distance is minimized. When the restrictions are placed on moment tensors $\mu_r(\varepsilon)$ this approach falls in the class of generalized moment estimation (GMM) [e.g. Hansen, 1982], see Hall [2005] for a textbook treatment. When the restrictions are placed on cumulant tensors no general framework exists, but a similar route as for moments can be followed.

It is interesting to note that minimum distance estimators are commonly adopted in the ICA literature using diagonal tensor restrictions and Euclidean distance to measure distance [e.g. Hyvärinen et al., 2001b, Chapter 11]. For instance, the JADE algorithm of Cardoso and Souloumiac [1993] solves a minimum distance problem that considers (after pre-whitening) cumulant restrictions on κ_4. In our set-up we follow the GMM literature and measure distance in a statistically meaningful way in order to get optimal efficiency of the associated estimator and based on the results of Section 5.2 we also consider non-diagonal tensor restrictions.

To set up our approach let A_0 denote the true A. We fix $\mathcal{V} = \mathcal{V}(\mathcal{I})$, with $\mathcal{I}$ defined by either Definition 5.1 or 5.8, and let $\pi_{\mathcal{V}}$ be defined as the orthogonal projection from $S^r(\mathbb{R}^d)$ to $\mathcal{V}^\perp$. Note that $\pi_{\mathcal{V}}(T)$ simply gives the coordinates $T_{\boldsymbol{i}}$ for $\boldsymbol{i} \in \mathcal{I}$, i.e. the set of zero restricted higher order tensor entries.

For $h_r(\varepsilon) \in \mathcal{V}$ we define the function

$$g(A) \;:=\; \mathrm{vec}_u(A \bullet h_2(Y) - I_d,\; \pi_{\mathcal{V}}(A \bullet h_r(Y))) \;\in\; \mathbb{R}^{\binom{d+1}{2}+|\mathcal{I}|}\,, \tag{16}$$

where vec_u is the vectorization that takes the unique entries of an element in $S^2(\mathbb{R}^d) \oplus \mathcal{V}^\perp$ and stacks them in a vector that has length $d_g = \binom{d+1}{2} + |\mathcal{I}|$. The sample equivalent of $g(A)$ is denoted by $g_n(A)$ and replaces h_j by $\hat{h}_j$, for $j = 2, r$, where $\hat{h}_j$ denotes either the sample moments, denoted by $\hat{\mu}_j$, or the jth order k-statistic, denoted by k_j, which are computed from a given sample $\{Y_s\}_{s=1}^n$. The computation of the sample moments $\hat{\mu}_j$ requires no explanation and for k-statistics we refer to McCullagh [2018, Chapter 4] as well as the supplementary material Section S3 where we provide explicit computational formulas.

The population and sample objective functions that we consider are given by

$$L_W(A) \;=\; \|g(A)\|_W^2 \quad \text{and} \quad \hat{L}_W(A) \;=\; \|\hat{g}_n(A)\|_W^2, \tag{17}$$

where W is an $d_g \times d_g$ positive definite weighting matrix, $\|v\|_W^2 = v'Wv$. The following result is clear.

Lemma 6.1. *Suppose that (1) holds with $h_2(\varepsilon) = I_d$ and $h_r(\varepsilon)$ is a tensor that satisfies the conditions in Theorem 5.3 or 5.10, then $L_W(A) = 0$ if and only if $A = QA_0$ for $Q \in \mathrm{SP}(d)$.*

Proof. We have $L_W(A) = 0$ if and only if $g(A) = 0$, which is equivalent $A \bullet h_2(Y) = I_d$ and $A \bullet h_r(Y) \in \mathcal{V}$. Since (1) holds, we also have $A_0 \bullet h_2(Y) = I_2$ and $A_0 \bullet h_r(Y) \in \mathcal{V}$. It follows that $A_0^{-1}A \in \mathrm{O}(d)$, or in other words, $A = QA_0$ for some $Q \in \mathrm{O}(d)$. Further,

$$A \bullet h_r(Y) = QA_0 \bullet h_r(Y) = Q \bullet h_r(\varepsilon) \in \mathcal{V},$$

which implies that $Q \in \mathcal{G}_T(\mathcal{V})$ and by Theorem 5.3 or 5.10 we have $\mathcal{G}_T(\mathcal{V}) = \mathrm{SP}(d)$. □

Given a sample $\{Y_s\}_{s=1}^n$, and a sequence of positive semidefinite matrices W_n we define the estimator

$$\widehat{A}_{W_n} \;:=\; \arg\min_{A \in \mathcal{A}} \hat{L}_{W_n}(A), \tag{18}$$

where $\mathcal{A} \subseteq GL(d)$ is fixed in advance and W_n is a weighting matrix that may depend on the sample. Here by $\arg\min_{A\in\mathcal{A}}$ we mean an arbitrarly chosen element from the set of minimizers of $\hat{L}_{W_n}(A)$.

6.1. **Consistency.** We can show that this class gives consistent estimates for the true A_0 up to sign and permutation. A possible set of conditions is as follows.

Proposition 6.2 (Consistency)**.** *Suppose that $\{Y_s\}_{s=1}^n$ is i.i.d from model (1) and (i) $h_r(\varepsilon)$ satisfies the conditions in Theorem 5.3 or 5.10, (ii) $\mathcal{A} \subset GL(d)$ is compact and $QA_0 \in \mathcal{A}$ for some $Q \in SP(d)$ (iii) $W_n \xrightarrow{p} W$ and W is positive definite, (iv) $\mathbb{E}\|Y_s\|^r < \infty$. Then $\widehat{A}_{W_n} \xrightarrow{p} QA_0$ as $n \to \infty$ for some $Q \in \mathrm{SP}(d)$.*

The proof is included in the supplementary material.

Condition (i) corresponds to the identification assumptions that were derived in the previous section. Condition (ii) imposes that the permutations QA_0 lie in some compact subset $\mathcal{A} \subset GL(d)$. This can be relaxed at the expense of a more involved derivation for the required uniform law of large numbers. Condition (iii) imposes that the weighting matrix is positive definite and we will determine an optimal choice for W below. The moment condition (iv) is necessary for applying the law of large numbers.

6.2. **Asymptotic normality.** The weighting matrix W_n can take different forms. In the ICA literature W_n is often taken as the identity matrix [e.g. Comon and Jutten, 2010, Chapter 5], but we will show that different choices for W_n yield, at least in theory, more efficient estimates provided that sufficient moments of Y exist. Specifically, when we take W_n such that it is consistent for the inverse of

$$\Sigma \;=\; \lim_{n\to\infty} \mathrm{var}(\sqrt{n}\hat{g}_n(QA_0)) \tag{19}$$

we can ensure that the resulting estimate $\widehat{A}_{W_n}$ achieves minimal variance in the class of generalized cumulant estimators (18).

Let $G(A) \in \mathbb{R}^{d_g \times d^2}$ be the Jacobian matrix representing the derivative of the function $g : \mathbb{R}^{d\times d} \to \mathbb{R}^{d_g}$ defined in (16). Here, defining the Jacobian we think about g as a map from $\mathbb{R}^{d^2}$ vectorizing A.

Proposition 6.3 (Asymptotic normality)**.** *Suppose that the conditions of Proposition 6.2 hold, (v) $QA_0 \in \text{int}(\mathcal{A})$ for some $Q \in SP(d)$, (vi) $\mathbb{E}\|Y_i\|^{2r} < \infty$, and denote by $G = G(QA_0)$. Then*

$$\sqrt{n}\text{vec}[\widehat{A}_{W_n} - QA_0] \xrightarrow{d} N(0, (G'WG)^{-1}G'W\Sigma WG(G'WG)^{-1}) \tag{20}$$

for some $Q \in \text{SP}(d)$, where Σ is given in (19). Moreover, for any $\widehat{\Sigma}_n \xrightarrow{p} \Sigma$ we have that

$$\sqrt{n}\text{vec}[\widehat{A}_{\widehat{\Sigma}_n^{-1}} - QA_0] \xrightarrow{d} N(0, S) \ .$$

for some $Q \in \text{SP}(d)$ and $S = (G'\Sigma^{-1}G)^{-1}$.

This result allows for an interesting comparison. For ICA models under full independence an efficient estimation method is developed in Chen and Bickel [2006]. When we relax the independence assumption, and instead only restrict higher order cumulant entries, the efficient estimator is given by $\widehat{A}_{\widehat{\Sigma}_n^{-1}}$. Here efficiency is understood in the sense that S is smaller (in the Löwner ordering) when compared to the variance in (20) for any W. Chamberlain [1987] shows that for moment restrictions the estimator $\widehat{A}_{\widehat{\Sigma}_n^{-1}}$ attains the semi-parametric efficiency bound in the class of non-parametric models characterized by restrictions $T_{\boldsymbol{i}} = 0$ for $\boldsymbol{i} \in \mathcal{I}$.

Implementing this estimator can be done in different ways. Proposition 6.2 shows that QA_0 can be consistently estimated regardless of the choice of weighting matrix. Given such first stage estimate, using say $W_n = I_{d_g}$, we can estimate Σ consistently (under the assumptions of Proposition 6.3). With this estimate we can compute $\widehat{A}_{\widehat{\Sigma}_n^{-1}}$ from (18). While this estimate is efficient, the procedure can obviously be iterated until convergence to avoid somewhat arbitrarily stopping at the first iteration, see Hansen and Lee [2021] for additional motivation for iterative moment estimators. Additionally, we may also consider $W_n = \widehat{\Sigma}_n(A)^{-1}$ as a weighting matrix, hence parametrizing the asymptotic variance estimate as a function of A, and minimize the objective function (18) using this weighting matrix [e.g. Hansen et al., 1996]. The methodology for estimating Σ and S, under both moment and cumulant restrictions, is discussed in the Appendix S5. In Section S4 we also discuss several hypothesis tests that allow us to test the higher order tensor restrictions.

7. Numerical illustration

In this section we evaluate the numerical performance of the minimum distance estimators introduced above. We simulate data from some of the

TABLE 1. DISTRIBUTIONS FOR ERRORS

Code	Name	Definition
$\mathcal{N}$	Gaussian	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)$
$t(\nu=5)$	Student's t	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{\left(-\frac{\nu+1}{2}\right)}$
SKU	Skewed Unimodal	$\frac{1}{5}\mathcal{N}(0,1)+\frac{1}{5}\mathcal{N}\left(\frac{1}{2},(\frac{2}{3})^2\right)+\frac{3}{5}\mathcal{N}\left(\frac{13}{12},(\frac{5}{9})^2\right)$
KU	Kurtotic Unimodal	$\frac{2}{3}\mathcal{N}(0,1)+\frac{1}{3}\mathcal{N}\left(0,(\frac{1}{10})^2\right)$
BM	Bimodal	$\frac{1}{2}\mathcal{N}\left(-1,(\frac{2}{3})^2\right)+\frac{1}{2}\mathcal{N}\left(1,(\frac{2}{3})^2\right)$
SBM	Separated Bimodal	$\frac{1}{2}\mathcal{N}\left(-\frac{3}{2},(\frac{1}{2})^2\right)+\frac{1}{2}\mathcal{N}\left(\frac{3}{2},(\frac{1}{2})^2\right)$
SKB	Skewed Bimodal	$\frac{3}{4}\mathcal{N}(0,1)+\frac{1}{4}\mathcal{N}\left(\frac{3}{2},(\frac{1}{3})^2\right)$
TRI	Trimodal	$\frac{9}{20}\mathcal{N}\left(-\frac{6}{5},(\frac{3}{5})^2\right)+\frac{9}{20}\mathcal{N}\left(\frac{6}{5},(\frac{3}{5})^2\right)+\frac{1}{10}\mathcal{N}\left(0,(\frac{1}{4})^2\right)$
CL	Claw	$\frac{1}{2}\mathcal{N}(0,1)+\sum_{l=0}^{4}\frac{1}{10}\mathcal{N}\left(l/2-1,(\frac{1}{10})^2\right)$
ACL	Asymmetric Claw	$\frac{1}{2}\mathcal{N}(0,1)+\sum_{l=-2}^{2}\frac{2^{1-l}}{31}\mathcal{N}\left(l+1/2,(2^{-l}/10)^2\right)$

Notes: The table reports the distributions that are used in the simulation studies to draw the errors. The mixture distributions are taken from Marron and Wand [1992], see their table 1.

non-independent components models of Section 2 and compare the estimation accuracy of the minimum distance estimators of Section 6 to several popular ICA methods.

7.1. **Common variance component models.** We start by simulating independent samples $\{Y^1,\ldots,Y^n\}$ from the common variance model (3), where $K=(1,\ldots,1)'$, $\tau\sim$ gamma$(1,1)$, ϕ is the identity function and $\eta\in\mathbb{R}^d$ has independent components that are simulated from different univariate distributions that are summarized in Table 1. The draws for η_i are standardized such that $\varepsilon_i=\tau\eta_i$ has mean zero and unit variance.

The matrix A is defined by $A=R'L$, where L is lower triangular with ones on the main diagonal and zeros elsewhere, and R is a rotation matrix that is parametrized by the Cayley transform of a skewed symmetric matrix who entries are randomly drawn for each sample from a $\mathcal{N}(0,I_l)$ distribution, with $l=d(d-1)/2$.

We compare the performance of several estimators. The minimum distance estimators from Section 6 are used with either $\kappa_3(\varepsilon)=\mu_3(\varepsilon)$ set to have zero off-diagonal elements, or $\mu_4(\varepsilon)$ restricted to be reflectionally invariant. We consider the weighting matrices $W_n=I_{d_g}$ and the asymptotically optimal choice $W_n=\widehat{\Sigma}_n^{-1}$.

TABLE 2. AMARI ERRORS: COMMON VARIANCE MODEL

Non-Independent Components Analysis										
Method	$\mathcal{N}$	$t(5)$	SKU	KU	BM	SBM	SKB	TRI	CL	ACL
$\mu_3^{d,I}$	0.45	0.35	0.33	0.35	0.53	0.65	0.46	0.56	0.46	0.43
$\mu_3^{d,\widehat{\Sigma}_n^{-1}}$	0.40	0.34	0.31	0.33	0.45	0.54	0.39	0.46	0.40	0.37
$\mu_4^{r,I}$	0.29	0.28	0.29	0.25	0.26	0.18	0.29	0.26	0.31	0.30
$\mu_4^{r,\widehat{\Sigma}_n^{-1}}$	0.30	0.28	0.30	0.25	0.24	0.12	0.27	0.22	0.28	0.29
TICA	0.37	0.19	0.27	0.05	0.80	0.91	0.70	0.83	0.53	0.46
Independent Components Analysis										
Method	$\mathcal{N}$	$t(5)$	SKU	KU	BM	SBM	SKB	TRI	CL	ACL
Fast	0.44	0.37	0.39	0.35	0.57	0.66	0.51	0.58	0.46	0.47
JADE	0.44	0.35	0.37	0.28	0.60	0.67	0.55	0.62	0.49	0.48
Kernel	0.45	0.36	0.39	0.36	0.55	0.64	0.49	0.56	0.46	0.46
ProDen	0.45	0.34	0.39	0.31	0.60	0.66	0.54	0.61	0.50	0.49
Efficient	0.44	0.48	0.46	0.48	0.40	0.44	0.40	0.42	0.41	0.41
NPML	0.44	0.42	0.43	0.43	0.45	0.42	0.44	0.43	0.43	0.42

Notes: The table reports the average Amari errors (across $S = 1000$ simulations) for data sampled from the common variance model (3) with $d = 2$ and $n = 200$. The columns correspond to the different errors considered for the components of η, see Table 1. The top panel reports the errors for the minimum distance methods and Topographical ICA (TICA). For the minimum distance methods we consider diagonal (d) and reflectionally invariant (r) restrictions for different order tensors μ_3, μ_4, combined with weighting matrices $W_n = I_d, \widehat{\Sigma}_n^{-1}$. The bottom panel reports comparison results for different independent component analysis methods: FastICA [Hyvärinen, 1999], JADE Cardoso and Souloumiac [1993], kernel ICA [Bach and Jordan, 2003], ProDenICA [Hastie and Tibshirani, 2002], efficient ICA [Chen and Bickel, 2006] and non-parametric ML ICA [Samworth and Yuan, 2012].

As an alternative non-independent component method we include topographical ICA (TICA) of Hyvärinen et al. [2001a]. We note that TICA assumes that mapping from the independent components that determines the variance to the errors is known, i.e. K is assumed known and explicitly used in the construction of the objective function [see Hyvärinen et al., 2001a, equation 3.10]. The minimum distance estimators that we propose do not exploit this knowledge.

For comparison purposes we also include FastICA [Hyvärinen, 1999], JADE [Cardoso and Souloumiac, 1993], kernel ICA [Bach and Jordan, 2003], ProDenICA [Hastie and Tibshirani, 2002], efficient ICA [Chen and Bickel, 2006] and non-parametric MLE ICA [Samworth and Yuan, 2012]. We stress that none of these alternative methods are designed for non-independent

components models and they are merely included to highlight that incorrectly imposing independence leads to distorted estimates.

For each simulation design we sample $S = 1000$ datasets and measure the accuracy of the estimates using the Amari et al. [1996] error:

$$d_A(\widehat{A}_{W_n}, A_0) = \frac{1}{2d}\sum_{j=1}^{d}\left(\frac{\sum_{j=1}^{n}|a_{ij}|}{\max_j |a_{ij}|} - 1\right) + \frac{1}{2d}\sum_{j=1}^{d}\left(\frac{\sum_{i=1}^{n}|a_{ij}|}{\max_i |a_{ij}|} - 1\right) ,$$

where a_{ij} is the i, j element of $A_0\widehat{A}_{W_n}^{-1}$. We report the averages of this error over the S datasets. In the supplementary material we also show results for the minimum distance index [Ilmonen et al., 2010].

Table 2 shows the baseline estimation results for $d = 2$ and $n = 200$. We find that minimum distance methods based on fourth order reflectionally invariant tensors always perform better when compared to the methods that rely on the independence assumption. The magnitude of the increase in the Amari errors differs across the different choices for the underlying densities. Notably with multi-modal densities the gains in estimation accuracy are large, often reducing the Amari error by more than half. Using the efficient weighting matrix is generally preferable when compared to the identity weighting, although the differences are not always large.

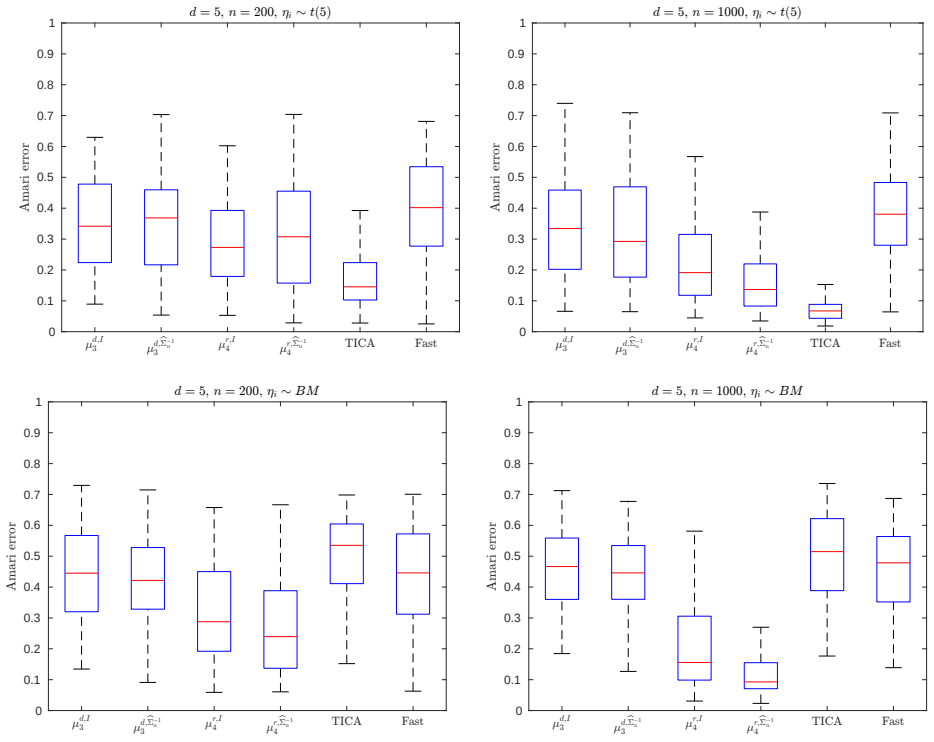
Minimum distance methods based on third order diagonal moment restrictions perform well when the true density has strong skewness (e.g. SKU). When used with the efficient weighting matrix this minimum distance method nearly always outperform the ICA methods, but even then the reflectionally invariant approach appears preferable.

As an alternative non-independent component method, TICA works well when the true likelihood is close to the imposed objective function of TICA. Generally, this is the case for all densities that are similar to the Student's t density. When the true density is far from the approximating densities TICA can have very large errors. A similar observation was made in the context of independent component analysis for pseudo maximum likelihood methods in Lee and Mesters [2021].

In Figure 1 we show that these findings persist for higher dimensional models and for large sample sizes. We show the results for $\eta_i \sim t(5)$ and $\eta_i \sim BM$. A larger selection of experiments is shown in the supplementary material. Two observations are worth pointing out. First, as the sample size n increases the Amari errors become smaller for the minimum distance methods, supporting the consistency result from Proposition 6.2. In contrast, for the ICA methods (here exemplified by FastICA) there is no change in accuracy when increasing n. Second, when the dimensions increase the ordering, in terms of performance, remains similar as above.

7.2. **Multiple scaled elliptical component models.** Next, we generate data from the nICA model with multiple scaled elliptical errors as in (4). Specifically, η is drawn from the uniform distribution on the d-sphere and τ

FIGURE 1. SELECTED COMMON VARIANCE EXPERIMENTS



Notes: The figure shows the boxplots for the Amari errors (across $S = 100$ simulations) for data sampled from the common variance model (3). The different settings for the simulations designs are described in the titles and the x-labels indicate the different estimation methods used.

is simulated as $\tau = Ke$, where K is a square matrix of ones and $e \in \mathbb{R}^d$ has independent components that are drawn from the distributions in Table (1). This implies that both the components of τ and η are dependent and the diagonal tensor identification result no longer holds. That said, Corollary 5.14 show that the reflectionally invariant moment tensors can still be used for identification. Also, in this setting topographical ICA cannot be used as the components of η are no longer independent. The other parts of the simulation design are similar as above and we perform the same comparisons.

The results are shown in Table 3. We find that the minimum distance methods based on the reflectionally invariant tensor restrictions now outperform all other methods. Using the efficient weighting matrix is not always preferable, as for most densities the weighting matrix is not estimated very accurately leading to more poor performance. The differences across the different densities for e are often small and the errors made by the ICA methods are quite similar.

TABLE 3. AMARI ERRORS: SCALED ELLIPTICAL

Non-Independent Components Analysis										
Method	$\mathcal{N}$	$t(5)$	SKU	KU	BM	SBM	SKB	TRI	CL	ACL
$\mu_3^{d,I}$	0.44	0.44	0.45	0.44	0.45	0.43	0.44	0.45	0.43	0.43
$\mu_3^{d,\widehat{\Sigma}_n^{-1}}$	0.43	0.41	0.42	0.42	0.44	0.43	0.43	0.44	0.41	0.41
$\mu_4^{r,I}$	0.21	0.25	0.23	0.23	0.21	0.21	0.22	0.25	0.24	0.25
$\mu_4^{r,\widehat{\Sigma}_n^{-1}}$	0.26	0.24	0.25	0.26	0.26	0.21	0.27	0.24	0.23	0.24
TICA	0.34	0.38	0.34	0.36	0.34	0.36	0.36	0.35	0.35	0.36

Independent Components Analysis										
Method	$\mathcal{N}$	$t(5)$	SKU	KU	BM	SBM	SKB	TRI	CL	ACL
Fast	0.43	0.43	0.44	0.43	0.42	0.44	0.44	0.44	0.44	0.46
JADE	0.45	0.43	0.44	0.44	0.44	0.43	0.45	0.45	0.45	0.44
Kernel	0.46	0.43	0.44	0.44	0.42	0.45	0.44	0.44	0.44	0.45
ProDen	0.45	0.43	0.44	0.45	0.42	0.45	0.45	0.45	0.44	0.46
Efficient	0.44	0.43	0.45	0.43	0.44	0.42	0.44	0.44	0.44	0.44
NPML	0.42	0.43	0.43	0.43	0.42	0.42	0.43	0.42	0.43	0.42

Notes: The table reports the average Amari errors (across $S = 1000$ simulations) for data sampled from the multiple scaled elliptical model (4) with $d = 2$ and $n = 200$. The columns correspond to the different errors considered for the components of η, see Table 1. The top panel reports the errors for the minimum distance methods and Topographical ICA (TICA). For the minimum distance methods we consider diagonal (d) and reflectionally invariant (r) restrictions for different order tensors μ_3, μ_4, combined with weighting matrices $W_n = I_d, \widehat{\Sigma}_n^{-1}$. The bottom panel reports comparison results for different independent component analysis methods: FastICA [Hyvärinen, 1999], JADE Cardoso and Souloumiac [1993], kernel ICA [Bach and Jordan, 2003], ProDenICA [Hastie and Tibshirani, 2002], efficient ICA [Chen and Bickel, 2006] and non-parametric ML ICA [Samworth and Yuan, 2012].

In the supplementary material Section S6 we provide a number of additional results. We show different error metrics, sample sizes and dimensions. The main conclusion — incorrectly imposing independence is costly — is found to hold across these variations

8. CONCLUSION

In the ICA literature identifiability of (1) is assured when ε has independent components out of which at most one is Gaussian. Although in the classical ICA literature independence seems a natural assumption, in many other applications it is considered too strong.

Our paper proposes a general framework to study weak conditions under which A in $AY = \varepsilon$ is identified up to the set of signed permutations. We

develop a novel approach to study this identifiability problem in the case of zero restrictions on fixed order moments or cumulants of ε. We obtain positive results for some useful zero patterns. These results can be used under strictly weaker conditions than independence and ensure the identifiability of several popular non-independent component models, e.g. common variance, scaled elliptical and mean independent components models.

While we have focused on relaxing the independence assumption in (1), it is easy to see that similar techniques can be used to relax independence assumptions in other linear models; e.g. measurement error models [Schennach, 2021], triangular systems [Lewbel et al., 2021], and structural vector autoregressive models [Kilian and Lütkepohl, 2017].

Acknowledgements

We would like to thank Joe Kileel, Mateusz Michałek, Mikkel Plagborg-Møller and Anna Seigal for helpful remarks.

References

Shun-ichi Amari, Andrzej Cichocki, and Howard Yang. A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems*, 8:757–763, 12 1996.

Manabu Asai, Michael McAleer, and Jun Yu. Multivariate stochastic volatility: A review. *Econometric Reviews*, 25(2-3):145–175, 2006.

Adelchi Azzalini and Marc G Genton. Robust likelihood methods based on the skew-t and related distributions. *International Statistical Review*, 76 (1):106–129, 2008.

Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 07 2002.

Francis R. Bach and Michael I. Jordan. Beyond independent components: Trees and clusters. *Journal of Machine Learning Research*, 4:1205–1233, 2003.

A. D. Back and A. S. Weigend. A first application of independent component analysis to extracting structure from stock returns. *International Journal of Neural Systems*, 8(4):473–484, 1997.

Geert Bekaert, Eric Engstrom, and Andrey Ermolov. Macro risks and the term structure of interest rates. *Journal of Financial Economics*, 141(2): 479–504, 2021.

Geert Bekaert, Eric Engstrom, and Andrey Ermolov. Identifying aggregate demand and supply shocks using sign restrictions and higher-order moments. 2022. working paper.

Jean-François Cardoso. Source separation using higher order moments. In *International Conference on Acoustics, Speech, and Signal Processing,*, pages 2109–2112 vol.4, 1989.

Jean-François Cardoso. High-Order Contrasts for Independent Component Analysis. *Neural Computation*, 11(1):157–192, 01 1999.

Jean-François Cardoso and A. Souloumiac. Blind beamforming for non-gaussian signals. *IEE Proceedings F - Radar and Signal Processing*, 140, 1993. doi: 10.1049/ip-f-2.1993.0054.

Gary Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987.

Aiyou Chen and Peter J. Bickel. Efficient independent component analysis. *The Annals of Statistics*, 34(6):2825 – 2855, 2006.

Pierre Comon. Independent component analysis, A new concept? *Signal Processing*, 36, 1994.

Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation*. Academic Press, Oxford, 2010.

Georges Darmois. Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 21(1/2): 2–8, 1953.

Richard Davis and Serena Ng. Time Series Estimation of the Dynamic Effects of Disaster-Type Shocks. Working paper, 2022.

Thorsten Drautzburg and Jonathan H Wright. Refining set-identification in vars through independence. Working Paper 29316, National Bureau of Economic Research, 2021.

Timothy Erickson, Colin Huan Jiang, and Toni M. Whited. Minimum distance estimation of the errors-in-variables model using linear cumulant equations. *Journal of Econometrics*, 183(2):211–221, 2014.

Jan Eriksson and Visa Koivunen. Identifiability and separability of linear ica models revisited. In *4th International Symposium on Independent Components Analysis and Blind Source Separation (ICA2003)*, volume 1, pages 23–27, 2003.

Florence Forbes and Darren Wraith. A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering. *Statistics and Computing*, 24(6):971–984, 2014.

Robert C. Geary. Inherent relations between random variables. *Proceedings of the Royal Irish Academy. Section A: Mathematical and Physical Sciences*, 47:63–76, 1941.

Alain Guay. Identification of structural vector autoregressions through higher unconditional moments. *Journal of Econometrics*, 225(1):27–46, 2021.

Alastair R. Hall. *Generalized Method of Moments*. Oxford University Press, 2005.

Marc Hallin and Chintan Mehta. R-estimation for asymmetric independent component analysis. *Journal of the American Statistical Association*, 110 (509):218–232, 2015.

Fang Han and Han Liu. ECA: High-dimensional elliptical component analysis in non-Gaussian distributions. *Journal of the American Statistical Association*, 113(521):252–268, 2018.

Bruce E. Hansen and Seojeong Lee. Inference for iterated gmm under misspecification. *Econometrica*, 89(3):1419–1447, 2021.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

Lars Peter Hansen, John Heaton, and Amir Yaron. Finite-Sample Properties of Some Alternative GMM Estimators. *Journal of Business & Economic Statistics*, 14(3):262–280, 1996.

Trevor Hastie and Rob Tibshirani. Independent components analysis through product density estimation. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, page 665–672, Cambridge, MA, USA, 2002. MIT Press.

Fumio Hayashi. *Econometrics*. Princeton University Press, 2000.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

Aapo Hyvärinen. Independent component analysis: recent advances. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110534, 2013. doi: 10.1098/rsta.2011.0534.

Aapo Hyvärinen, Patrik O Hoyer, and Mika Inki. Topographic independent component analysis. *Neural computation*, 13(7):1527–1558, 2001a.

Aapo Hyvärinen, Juha Karhunen, and Errki Oja. *Independent Component Analysis*. Wiley, New York., 2001b.

Pauliina Ilmonen and Davy Paindaveine. Semiparametrically efficient inference based on signed ranks in symmetric independent component models. *The Annals of Statistics*, 39(5):2448 – 2476, 2011.

Pauliina Ilmonen, Klaus Nordhausen, Hannu Oja, and Esa Ollila. A new performance index for ICA: Properties, computation and asymptotic analysis. In Vincent Vigneron, Vicente Zarzoso, Eric Moreau, Rémi Gribonval, and Emmanuel Vincent, editors, *Latent Variable Analysis and Signal Separation*, pages 229–236, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

Douglas Kelker. Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 419–430, 1970.

Lutz Kilian and Helmut Lütkepohl. *Structural Vector Autoregressive Analysis*. Cambridge University Press, 2017.

Markku Lanne and Jani Luoto. Gmm estimation of non-gaussian structural vector autoregression. *Journal of Business & Economic Statistics*, 39(1): 69–81, 2021.

Adam Lee and Geert Mesters. Robust non-gaussian identification and inference for simultaneous equations. *Working Paper*, 2021.

Arthur Lewbel, Susanne M. Schennach, and Linqi Zhang. Identification of a triangular two equation system without instruments. 2021. working paper.

Lek-Heng Lim. Tensors in computations. *Acta Numerica*, 30:555–764, 2021.

Sydney C. Ludvigson, Sai Ma, and Serena Ng. Uncertainty and business cycles: Exogenous impulse or endogenous response? *American Economic Journal: Macroeconomics*, 13(4), 2021.

Eugene Lukacs. Some extensions of a theorem of Marcinkiewicz. *Pacific Journal of Mathematics*, 8(3):487–501, 1958.

Józef Marcinkiewicz. Sur une propriété de la loi de Gauss. *Mathematische Zeitschrift*, 44(1):612–618, 1939.

J Steve Marron and Matt P Wand. Exact mean integrated squared error. *The Annals of Statistics*, pages 712–736, 1992.

David S. Matteson and Ruey S. Tsay. Independent component analysis via distance covariance. *Journal of the American Statistical Association*, 112 (518):623–637, 2017.

Peter McCullagh. *Tensor methods in statistics: Monographs on statistics and applied probability*. Chapman and Hall/CRC, 2018.

A Meyer-Base, Ddrothee Auer, and Axel Wismueller. Topographic independent component analysis for fmri signal detection. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 1, pages 601–605. IEEE, 2003.

Anke Meyer-Bäse, Oliver Lange, Axel Wismüller, and Helge Ritter. Model-free functional mri analysis using topographic independent component analysis. *International journal of neural systems*, 14(04):217–228, 2004.

José Luis Montiel Olea, Mikkel Plagborg-Møller, and Eric Qian. Svar identification from higher moments: Has the simultaneous causality problem been solved? *AEA Papers and Proceedings*, 112:481–85, May 2022.

Jason A Palmer, Ken Kreutz-Delgado, Bhaskar D Rao, and Scott Makeig. Modeling and estimation of dependent subspaces with non-radially symmetric and skewed densities. In *Independent Component Analysis and Signal Separation: 7th International Conference, ICA 2007, London, UK, September 9-12, 2007. Proceedings 7*, pages 97–104. Springer, 2007.

David Rossell and Piotr Zwiernik. Dependence in elliptical partial correlation graphs. *Electronic Journal of Statistics*, 15(2):4236–4263, 2021.

Richard J. Samworth and Ming Yuan. Independent component analysis via nonparametric maximum likelihood estimation. *The Annals of Statistics*, 40(6):2973 – 3002, 2012.

Susanne M. Schennach. Measurement systems. *Journal of Economic Literature*, 2021. forthcoming.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006.

V. P. Skitovic. On a property of the normal distribution. *Dokl. Akad. Nauk SSSR (N.S.)*, 89:217–219, 1953.

Terry P. Speed. Cumulants and partition lattices 1. *Australian Journal of Statistics*, 25(2):378–388, 1983.

Carlos Velasco. Identification and estimation of structural varma models using higher order dynamics. *Journal of Business & Economic Statistics*, 2022. forthcoming.

Daniel Vogel and Roland Fried. Elliptical graphical modelling. *Biometrika*, 98(4):935–951, 2011.

Piotr Zwiernik. L-cumulants, L-cumulant embeddings and algebraic statistics. *Journal of Algebraic Statistics*, 3(1):11 – 43, 2012.

Piotr Zwiernik. Semialgebraic statistics and latent tree models. *Monographs on Statistics and Applied Probability*, 146:146, 2016.

Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain

Email address: geert.mesters@upf.edu

Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada

Email address: piotr.zwiernik@utoronto.ca