

# NON-INDEPENDENT COMPONENTS ANALYSIS

BY GEERT MESTERS<sup>1,a</sup> , PIOTR ZWIERNIK<sup>2,b</sup> 

<sup>1</sup>*Department of Economics and Business, Universitat Pompeu Fabra, [geert.mesters@upf.edu](mailto:geert.mesters@upf.edu)*

<sup>2</sup>*Department of Statistical Sciences, University of Toronto, [piotr.zwiernik@utoronto.ca](mailto:piotr.zwiernik@utoronto.ca)*

A seminal result in the ICA literature states that for  $AY = \varepsilon$ , if the components of  $\varepsilon$  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  $\varepsilon$ . 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

$$(1) \quad AY = \varepsilon,$$

where  $Y \in \mathbb{R}^d$  is observed,  $A \in \mathbb{R}^{d \times d}$  is invertible, and  $\varepsilon$  is a mean-zero hidden random vector with uncorrelated components. If  $\varepsilon$  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  $\varepsilon$  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, Karhunen and Oja, 2001; Comon and Jutten, 2010; Hyvärinen, 2013).

As implied by its name, the working assumption in the ICA literature is that the components of  $\varepsilon$  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, Karhunen and Oja (2001, 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, Hoyer and Inki, 2001; Matteson and Tsay, 2017).

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  $\Omega = \{QA : Q \in O(d)\}$ , where  $O(d)$  is the set of  $d$ -dimensional orthogonal matrices. This refinement allows to formally state our research question: Which higher order restrictions on  $\varepsilon$  allow to identify a finite, possibly structured, subset of  $\Omega$ ?

We systematically study our question by considering different restrictions on the higher order moments or cumulants of  $\varepsilon$ . We focus on cases where a subset of entries of a given  $r$ th order moment/cumulant tensor are set to zero, for some  $r > 2$ . Although there are alternative

---

*MSC2020 subject classifications:* Primary 15A69, 62H99; secondary .

*Keywords and phrases:* independent component analysis, identifiability, cumulants, tensors.

types of restrictions that can be considered, zero restrictions are attractive as they can often be justified by generative models for  $\varepsilon$  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, Engstrom and Ermolov \(2021\)](#) 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. For instance the JADE algorithm of [Cardoso and Souloumiac \(1993\)](#) is based on diagonalizing the fourth order cumulant tensor. We show that, without imposing the independence assumption, if we set the off-diagonal elements of *any*  $r$ th 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  $\varepsilon$  are not independent. For instance, if  $\varepsilon$  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, Hoyer and Inki, 2001](#); [Montiel Olea, Plagborg-Møller and Qian, 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.

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 exist — 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, 1999a; 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 Imonen 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, Jiang and Whited (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  $\varepsilon$  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

$$(2) \quad AY = \varepsilon, \quad \text{with} \quad \varepsilon = \tau\eta,$$

where  $\tau$  is some positive random variable with finite second moment and  $\eta$  is a random vector that is independent of  $\tau$  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  $\varepsilon$  remain uncorrelated. However, even if the components of  $\eta$  are independent, the components of  $\varepsilon$  are generally not. Indeed, assuming  $\eta$  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, Hoyer and Inki (2001), which can be used in image analysis Meyer-Base, Auer and Wismueller (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, McAleer and Yu (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, Ma and Ng, 2021](#)).

The non-independence for the baseline common variance model (2) carries over to more general models, where  $\tau$  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

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

where  $\eta \in \mathbb{R}^d$  and  $Z \in \mathbb{R}^m$  are independent random vectors with independent components, the function  $\phi: \mathbb{R} \rightarrow \mathbb{R}_{>0} = \{x \in \mathbb{R} : x > 0\}$ , is applied coordinatewise and  $\odot$  denotes the Hadamard product. In this model  $\tau \in \mathbb{R}^m$  is independent of  $\eta$ , and the components of  $\varepsilon$  share a common variance if they load on the same underlying components, or factors,  $Z$ . By exactly the same argument as above  $\varepsilon$  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  $\varepsilon$  is said to follow an elliptical distribution ([Kelker, 1970](#)). It follows that, generally,  $\varepsilon$  will not have independent components. The exception is the case where  $\tau^2$  follows a  $\chi_d^2$  distribution, such that  $\varepsilon$  is standard normal.

In general, for elliptical errors  $A$  can never be recovered beyond the set  $\Omega = \{QA : Q \in O(d)\}$ . 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  $\tau$  a  $d$ -dimensional vector. Such *multiple scale elliptical* distribution continues to have non-independent components but any variation in the components of  $\tau$  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

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

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  $A = (a_1, \dots, a_d)'$ . 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  $\varepsilon$  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

$$(5) \quad a'_i Y = \varepsilon_i, \quad \text{with} \quad \mathbb{E}(\varepsilon_i | \varepsilon_{-i}) = 0, \quad \text{for } i = 1, \dots, d,$$

where  $\varepsilon_{-i}$  drops  $\varepsilon_i$  from  $\varepsilon$ . This relaxed notion of independence is attractive in practice as it avoids restricting terms like  $\mathbb{E}(\varepsilon_i^k | \varepsilon_{-i})$  for  $k = 2, 3, \dots$ , 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  $\varepsilon_1$  and  $\varepsilon_2$  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  $k = 2, 3, \dots$  are typically not motivated by economic theory. Additional economic motivation for specific moment conditions in supply and demand models is given in [Bekaert, Engstrom and Ermolov \(2021, 2022\)](#).

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

$$\mu_r(X)_{i_1 \dots i_r} = \mathbb{E}X_{i_1} \dots X_{i_r} = \left. \frac{\partial^r}{\partial t_{i_1} \dots \partial t_{i_r}} M_X(\mathbf{t}) \right|_{\mathbf{t}=0}.$$

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

$$\kappa_r(X)_{i_1 \dots i_r} = \text{cum}(X_{i_1}, \dots, X_{i_r}) = \left. \frac{\partial^r}{\partial t_{i_1} \dots \partial t_{i_r}} K_X(\mathbf{t}) \right|_{\mathbf{t}=0}.$$

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  $d \times \dots \times d$  order  $r$  tensors is denoted by  $S^r(\mathbb{R}^d)$ . Writing  $[d] = \{1, \dots, d\}$ , the set of indices of an order  $r$  tensor is  $[d]^r$ . However,  $S^r(\mathbb{R}^d) \subset \mathbb{R}^{d \times \dots \times d}$  has dimension  $\binom{d+r-1}{r}$  and the unique entries of  $T \in S^r(\mathbb{R}^d)$  are  $T_{i_1 \dots i_r}$  for  $1 \leq i_1 \leq \dots \leq i_r \leq d$ .

The vast majority of results in this paper holds for both moment and cumulant tensors. To avoid excessive notation we denote a given  $r$ th 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

$$(6) \quad h_r(AX) = A \bullet h_r(X),$$

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

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

for all  $(i_1, \dots, i_r) \in [d]^r$ , 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  $\mathbb{R}^{n_1 \times \dots \times n_r} \rightarrow \mathbb{R}^{m_1 \times \dots \times m_r}$  given by matrices  $A^{(1)} \in \mathbb{R}^{m_1 \times n_1}, \dots, A^{(r)} \in \mathbb{R}^{m_r \times n_r}$ :

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

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  $\text{var}(\varepsilon) = I_d$ , the variance of  $Y$  satisfies  $\text{var}(Y) = (A^T A)^{-1}$  and so it is enough to narrow down potential candidates for  $A$  to the compact set

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

Our main insight is as follows: Since  $\varepsilon$  is unobserved, multiplying (1) by  $Q \in O(d)$  gives an alternative representation  $\tilde{A}Y = \tilde{\varepsilon}$ , where  $\mathbb{E}\tilde{\varepsilon} = 0$  and  $\text{var}(\tilde{\varepsilon}) = I_d$ . The goal is to define suitable additional restrictions on the distribution of  $\varepsilon$  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

$$(8) \quad T = h_r(AY) = A \bullet h_r(Y) ,$$

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

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

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

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

which is the subset of  $\Omega$  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  $\text{var}(\varepsilon) = I_d$ . 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*

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

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  $SP(d)$ . These are the  $2^d d!$  matrices that are of the form  $DP$ , where  $D, P \in O(d)$  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  $\kappa_r(\varepsilon) = \text{cum}_r(\varepsilon)$  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  $(i_1, \dots, i_r)$  satisfying  $1 \leq i_1 \leq \dots \leq i_r \leq d$  and by defining the vector space  $\mathcal{V} = \mathcal{V}(\mathcal{I})$  of symmetric tensors  $T \in S^r(\mathbb{R}^d)$  such that  $T_{\mathbf{i}} = 0$  for all  $\mathbf{i} = (i_1, \dots, i_r) \in \mathcal{I}$ . In symbols:

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

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

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  $\mathcal{G}_T(\mathcal{V}) = \text{SP}(2)$ .

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  $\mathbb{E}\varepsilon = 0$ ,  $\text{var}(\varepsilon) = I_d$  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  $r$ th order moment or cumulant tensor of  $\varepsilon$ . 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  $T_{\mathbf{i}} = 0$  unless  $\mathbf{i} = (i, \dots, i)$  for some  $i = 1, \dots, d$ . We define  $\mathcal{V}^{\text{diag}}$  as the set of diagonal tensors in  $S^r(\mathbb{R}^d)$ .

Of course, if the components of  $\varepsilon$  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  $\varepsilon$  are zero, without imposing any restrictions on the higher order moments.

For verifying whether  $\mathcal{V}^{\text{diag}}$  provides sufficient identifying restrictions we will study the tensors  $T$  and  $Q \bullet T$  via their associated homogeneous polynomials in variables  $x = (x_1, \dots, x_d)$ . We have

$$(12) \quad f_T(x) = \sum_{i_1=1}^d \cdots \sum_{i_r=1}^d T_{i_1 \dots i_r} x_{i_1} \cdots x_{i_r} = \sum_{\mathbf{i}} T_{\mathbf{i}} (x^{\otimes r})_{\mathbf{i}} = \langle T, x^{\otimes r} \rangle,$$

where  $x^{\otimes r} \in S^r(\mathbb{R}^d)$  denotes the tensor with coordinates  $(x^{\otimes r})_{i_1 \dots i_r} = x_{i_1} \cdots x_{i_r}$ . 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 satisfying*

$$(13) \quad T_{i\dots i} \neq 0 \quad \text{for at least } d-1 \text{ different values of } i = 1, \dots, d .$$

*Then  $Q \bullet T \in \mathcal{V}^{\text{diag}}$  if and only if  $Q \in \text{SP}(d)$ , i.e.  $\mathcal{G}_T(\mathcal{V}^{\text{diag}}) = \text{SP}(d)$ . If  $r = 2$ , the same conclusion holds under a stronger condition that  $T_{jj} \neq T_{kk}$  for  $j \neq k$ .*

**PROOF.** The left direction is straightforward. 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  $Q \in \text{SP}(d)$  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

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

If  $r \geq 3$ , this equality can hold only if  $T_{j\dots j} = T_{k\dots k} = 0$ , which is impossible by the genericity condition (13). If  $r = 2$  this cannot hold by the stronger condition  $T_{jj} \neq T_{kk}$  for  $j \neq k$ .  $\square$

**REMARK 5.4.** The genericity condition is a necessary condition. Indeed, if, for example  $T_{1\dots 1} = T_{2\dots 2} = 0$  then  $\frac{\partial}{\partial x_1^2} f_T(x) = \frac{\partial}{\partial x_2^2} f_T(x) = 0$ . 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  $\mathbb{E}\varepsilon = 0$ ,  $\text{var}(\varepsilon) = I_d$  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  $\varepsilon$  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  $\text{var}(\varepsilon) = I_d$ . Suppose that additionally  $\varepsilon$  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  $\phi: \mathbb{R} \rightarrow \mathbb{R}_{>0}$ , is applied coordinatewise.
- (b)  $\mathbb{E}(\varepsilon_i | \varepsilon_{-i}) = 0$  for  $i = 1, \dots, d$ .

Then  $h_3(\varepsilon) = \mu_3(\varepsilon) = \kappa_3(\varepsilon)$  is a diagonal tensor. If additionally, the genericity condition (13) holds 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.  $\square$

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  $T_{i_1 \dots i_r}$ , 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. We define  $\mathcal{V}^{\text{refl}}$  as the set of reflectionally invariant tensors in  $S^r(\mathbb{R}^d)$ .

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 follows 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  $D$  with  $\pm 1$  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  $D$  with  $\pm 1$  on the diagonal. 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.  $\square$

**THEOREM 5.10.** *Suppose that  $T \in S^r(\mathbb{R}^d)$  for an even  $r$  is a reflectionally invariant tensor. Let  $l = (r - 2)/2$  and suppose, in addition, that  $T$  satisfies*

$$(14) \quad \sum_{i_1, \dots, i_l} T_{i_1 i_1 \dots i_l i_l j j} \neq \sum_{i_1, \dots, i_l} T_{i_1 i_1 \dots i_l i_l k k} \quad \text{for all } j \neq k.$$

*Then  $Q \bullet T \in \mathcal{V}^{\text{refl}}$  if and only if  $Q \in \text{SP}(d)$ , i.e.  $\mathcal{G}_T(\mathcal{V}^{\text{refl}}) = \text{SP}(d)$ .*

**REMARK 5.11.** We emphasize that the genericity condition in (14) 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 (14) 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.

**REMARK 5.12.** If  $r = 2$  then  $\mathcal{V}^{\text{refl}} = \mathcal{V}^{\text{diag}}$  and it corresponds to the set of diagonal matrices. In this case, the genericity condition (14) and the genericity condition in Theorem 5.3 coincide:  $T_{jj} \neq T_{kk}$  for  $j \neq k$ .

Theorem 5.10 is proven using the following lemma.

**LEMMA 5.13.** *Let  $r$  be even and suppose that  $T \in S^r(\mathbb{R}^d)$  is a reflectionally invariant tensor satisfying (14). Then  $Q \bullet T = T$  for  $Q \in \text{O}(d)$  if and only if  $Q$  is a diagonal matrix with  $\pm 1$  on the main diagonal.*

**PROOF.** For the left implication assume that  $Q$  is diagonal with  $\pm 1$  on the diagonal. Then  $f_{Q \bullet T}(x) = f_T(Q'x) = f_T(x)$ , where the first equality follows from Lemma 5.2 and the second from Lemma 5.9 as  $T$  is reflectionally invariant. By result  $Q \bullet T = T$ .

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  $1 \leq i \leq j \leq d$

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

By the genericity condition (14), 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  $2 \times 2$  submatrix  $Q_{ij,ij}$  are not constrained, so  $Q$  does not have to be diagonal.

Suppose now that the claim is true for  $r \geq 2$  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  $Q \bullet T = (Q, \dots, Q) \cdot T$ . Now note that since  $Q \bullet T = T$  we can replace any  $Q$  by  $I_d$ . We have  $(Q, \dots, Q, I_d, I_d) \cdot T = T$  where

we replaced the last two  $Q$ 's by  $I_d$ . Now acting on both sides with  $(I_d, \dots, I_d, Q', Q')$  and using  $Q'Q = I_d$  gives

$$(15) \quad (Q, \dots, Q, I_d, I_d) \cdot T = (I_d, \dots, I_d, Q', Q') \cdot T.$$

We want to show that this equality implies that  $Q$  is a diagonal matrix with  $\pm 1$  on the main diagonal. Let  $\mathbf{i} = (i_1, \dots, i_r)$  and consider all  $(r+2)$ -tuples  $(\mathbf{i}, u, u)$  for some  $u \in \{1, \dots, d\}$ . Writing (15) restricted to these indices gives

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

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  $S_{\mathbf{i}} = \sum_u T_{i_1 \dots i_r u u}$ , summation over  $u$  yields

$$\sum_{j_1, \dots, j_r} Q_{i_1 j_1} \cdots Q_{i_r j_r} S_{j_1 \dots j_r} = \sum_v T_{i_1 \dots i_r v v} = S_{i_1 \dots i_r}.$$

Since this equation holds for every  $\mathbf{i} = (i_1, \dots, i_r)$ , we conclude  $Q \bullet S = S$ , where  $S \in S^r(\mathbb{R}^d)$  with entries  $S_{\mathbf{i}} = S_{i_1 \dots i_r}$ .

Note however that  $S$  is a reflectionally invariant tensor. Indeed, if some index appears in  $\mathbf{i}$  odd number of times then  $T_{i_1 \dots i_r} = 0$  for all  $u = 1, \dots, d$  as  $T$  is reflectionally invariant. By definition,  $S_{\mathbf{i}}$  is then also zero.

Since  $T$  satisfies (14),  $S$  satisfies (14) too. Indeed, for  $l = (r-2)/2$  we have

$$\sum_{i_1, \dots, i_l} S_{i_1 i_1 \dots i_l i_l j j} = \sum_{i_1, \dots, i_l} \sum_{i_{l+1}} T_{i_1 i_1 \dots i_l i_l i_{l+1} i_{l+1} j j}$$

and so these quantities are distinct for all  $j = 1, \dots, d$  by assumption on  $T$ .

Now, by the induction assumption, we conclude that  $Q$  is diagonal.  $\square$

**PROOF OF THEOREM 5.10.** The left implication is straightforward. For the right implication, suppose  $Q \in O(d)$  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  $D \in O(d)$ , 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 5.13 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  $i = 1, \dots, d$ , we conclude that  $Q \in \text{SP}(d)$ .  $\square$

Combining Proposition 4.1 and Theorem 5.10 implies the following result.

**THEOREM 5.14.** *Consider the model (1) with  $\mathbb{E}\varepsilon = 0$ ,  $\text{var}(\varepsilon) = I_d$  and suppose that for some even  $r$  the tensor  $h_r(\varepsilon)$  is reflectionally invariant and it satisfies the genericity condition (14). Then  $A$  is identifiable up to permuting and swapping signs of its rows.*

Using Theorem 5.14 we can provide identification results for all the models from Section 2. We summarize these all in the following statement.

**COROLLARY 5.15.** *Consider the model  $AY = \varepsilon$  with  $\mathbb{E}(\varepsilon) = 0$  and  $\text{var}(\varepsilon) = I_d$ . If  $\varepsilon$  follows the general common covariance model (3) then  $h_4(\varepsilon)$  is reflectionally invariant. If  $\varepsilon$  follows the multiple scaled elliptical distribution (4), then  $h_r(\varepsilon)$  is reflectionally invariant for every even  $r \geq 4$ . If  $\varepsilon$  is mean independent as in (5) then  $h_4(\varepsilon)$  is reflectionally invariant. If additionally, the genericity condition (14) 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

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

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

$$(16) \quad D_i(\tau \odot U) = \tau \odot (D_i U) \stackrel{d}{=} \tau \odot U,$$

where  $U, \tau$  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.  $\square$

Note that if  $\varepsilon$  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 (14). 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  $\sum_i T_{iijj} = d + 2$  for all  $i = 1, \dots, d$ .

**REMARK 5.16.** The genericity condition (14) can be worked out explicitly for each model. For instance, for the general common variance model (a) we have for  $\mu_4(\varepsilon)$  that

$$\sum_{l=1}^d \mathbb{E}(\tau_l^2 \tau_j^2) \mathbb{E}(\eta_l^2 \eta_j^2) \neq \sum_{l=1}^d \mathbb{E}(\tau_l^2 \tau_k^2) \mathbb{E}(\eta_l^2 \eta_k^2) \quad \text{for all } j \neq k.$$

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

$$2\mathbb{E}\tau_j^4 + \sum_{l \neq j} \mathbb{E}(\tau_j^2 \tau_l^2) \neq 2\mathbb{E}\tau_k^4 + \sum_{l \neq k} \mathbb{E}(\tau_k^2 \tau_l^2) \quad \text{for all } j \neq k.$$

**5.3. Generalizations.** Theorems 5.5 and 5.14 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  $\dim(\mathcal{O}(d)) = \binom{d}{2}$ , 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

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

implies that  $\mathcal{G}_T$  is finite, but in general  $\mathcal{G}_T \neq \text{SP}(d)$ . Example S8 explicitly computes the difference between  $\mathcal{G}_T$  and  $\text{SP}(d)$  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

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

CONJECTURE 5.17. If  $T$  is a generic tensor in  $\mathcal{V}(\mathcal{I})$  then  $Q \bullet T \in \mathcal{V}(\mathcal{I})$  if and only if  $Q \in \text{SP}(d)$ .

The case when  $d = 2$  is very special because  $\text{O}(2)$  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 5.18. *Suppose that  $T \in S^r(\mathbb{R}^2)$  satisfies  $T_{12\dots 2} = T_{1\dots 12} = 0$  but is otherwise generic. Then  $Q \bullet T \in \mathcal{V}(\mathcal{I})$  if and only if  $Q \in \text{SP}(2)$ .*

We prove this result in Supplement S0. The genericity conditions are again linear and can be recovered from the proof.

REMARK 5.19. Consider the two-dimensional supply and demand model described in Section 2.3. Proposition 5.18 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, Hoyer and Inki, 2001). 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, Karhunen and Oja, 2001, 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  $\kappa_4$ . 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 take  $\mathcal{V}$  as  $\mathcal{V}^{\text{diag}}$  or  $\mathcal{V}^{\text{refl}}$  as given in Definitions 5.1 and 5.8, respectively, 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  $i$  for which  $T_i = 0$ , i.e. the set of zero restricted higher order tensor entries. For diagonal tensors this is the set of all off-diagonal elements, while for reflectionally invariant tensors this is the set where at least one index appears an uneven number of times.

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

$$(17) \quad g(A) := \text{vec}_u(A \bullet h_2(Y) - I_d, \pi_{\mathcal{V}}(A \bullet h_r(Y))) \in \mathbb{R}^{\binom{d+1}{2} + |\mathcal{I}|},$$

where  $\text{vec}_u$  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 given by

$$(18) \quad \hat{g}_n(A) := \text{vec}_u(A \bullet \hat{h}_2 - I_d, \pi_{\mathcal{V}}(A \bullet \hat{h}_r)) \in \mathbb{R}^{\binom{d+1}{2} + |\mathcal{I}|},$$

where  $\hat{h}_j$  denotes either the sample moments, denoted by  $\hat{\mu}_j$ , or the  $j$ th 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 supplement S3 where we provide explicit computational formulas.

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

$$(19) \quad L_W(A) = \|g(A)\|_W^2 \quad \text{and} \quad \hat{L}_W(A) = \|\hat{g}_n(A)\|_W^2,$$

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

**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  $Q \in \text{SP}(d)$ .*

The lemma implies that the identifying results from Section 5 can be translated into an minimization problem. Given a sample  $\{Y_s\}_{s=1}^n$ , and a sequence of positive semidefinite matrices  $W_n$  we define the estimator

$$(20) \quad \hat{A}_{W_n} := \arg \min_{A \in \mathcal{A}} \hat{L}_{W_n}(A),$$

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 arbitrarily 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  $\hat{A}_{W_n} \xrightarrow{p} QA_0$  as  $n \rightarrow \infty$  for some  $Q \in SP(d)$ .*

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

$$(21) \quad \Sigma = \lim_{n \rightarrow \infty} \text{var}(\sqrt{n}\hat{g}_n(QA_0))$$

we can ensure that the resulting estimate  $\hat{A}_{W_n}$  achieves minimal variance in the class of generalized cumulant estimators (20).

Let  $G(A) \in \mathbb{R}^{d_g \times d^2}$  be the Jacobian matrix representing the derivative of the function  $g: \mathbb{R}^{d \times d} \rightarrow \mathbb{R}^{d_g}$  defined in (17). 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*

$$(22) \quad \sqrt{n} \text{vec}[\hat{A}_{W_n} - QA_0] \xrightarrow{d} N(0, (G'WG)^{-1}G'W\Sigma WG(G'WG)^{-1})$$

for some  $Q \in SP(d)$ , where  $\Sigma$  is given in (21). Moreover, for any  $\hat{\Sigma}_n \xrightarrow{p} \Sigma$  we have that

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

for some  $Q \in SP(d)$  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  $\hat{A}_{\hat{\Sigma}_n}$ . Here efficiency is understood in the sense that  $S$  is smaller (in the Löwner ordering) when compared to the variance in (22) for any  $W$ . Chamberlain (1987) shows that for moment restrictions the estimator  $\hat{A}_{\hat{\Sigma}_n}$  attains the semi-parametric efficiency bound in the class of non-parametric models characterized by restrictions  $T_i = 0$  for  $i \in \mathcal{I}$ .



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 $t$	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
SKU	Skewed Unimodal	$\frac{1}{5}\mathcal{N}(0, 1) + \frac{1}{5}\mathcal{N}\left(\frac{1}{2}, \left(\frac{2}{3}\right)^2\right) + \frac{3}{5}\mathcal{N}\left(\frac{13}{12}, \left(\frac{5}{9}\right)^2\right)$
KU	Kurtotic Unimodal	$\frac{2}{3}\mathcal{N}(0, 1) + \frac{1}{3}\mathcal{N}\left(0, \left(\frac{1}{10}\right)^2\right)$
BM	Bimodal	$\frac{1}{2}\mathcal{N}\left(-1, \left(\frac{2}{3}\right)^2\right) + \frac{1}{2}\mathcal{N}\left(1, \left(\frac{2}{3}\right)^2\right)$
SBM	Separated Bimodal	$\frac{1}{2}\mathcal{N}\left(-\frac{3}{2}, \left(\frac{1}{2}\right)^2\right) + \frac{1}{2}\mathcal{N}\left(\frac{3}{2}, \left(\frac{1}{2}\right)^2\right)$
SKB	Skewed Bimodal	$\frac{3}{4}\mathcal{N}(0, 1) + \frac{1}{4}\mathcal{N}\left(\frac{3}{2}, \left(\frac{1}{3}\right)^2\right)$
TRI	Trimodal	$\frac{9}{20}\mathcal{N}\left(-\frac{6}{5}, \left(\frac{3}{5}\right)^2\right) + \frac{9}{20}\mathcal{N}\left(\frac{6}{5}, \left(\frac{3}{5}\right)^2\right) + \frac{1}{10}\mathcal{N}\left(0, \left(\frac{1}{4}\right)^2\right)$
CL	Claw	$\frac{1}{2}\mathcal{N}(0, 1) + \sum_{l=0}^4 \frac{1}{10}\mathcal{N}(l/2 - 1, \left(\frac{1}{10}\right)^2)$
ACL	Asymmetric Claw	$\frac{1}{2}\mathcal{N}(0, 1) + \sum_{l=-2}^2 \frac{2^{1-l}}{31}\mathcal{N}(l + 1/2, (2^{-l}/10)^2)$

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.

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  $\Sigma$  consistently (under the assumptions of Proposition 6.3). With this estimate we can compute  $\hat{A}_{\hat{\Sigma}_n^{-1}}$  from (20). 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  $W_n = \hat{\Sigma}_n(A)^{-1}$  as a weighting matrix, hence parametrizing the asymptotic variance estimate as a function of  $A$ , and minimize the objective function (20) using this weighting matrix (e.g. [Hansen, Heaton and Yaron, 1996](#)). The methodology for estimating  $\Sigma$  and  $S$ , under both moment and cumulant restrictions, is discussed in the supplement S5. For moment tensors we recommend using a standard plug-in estimator, but for cumulant tensors it often more convenient to adopt a simple bootstrap that resamples the residuals to estimate  $\Sigma$  as the analytical form of  $\Sigma$  depends on Jacobian of the transformation of cumulants to moments which can be tedious to derive in practice. This residual bootstrap is valid under the assumptions Proposition 6.3, see supplement S5. In supplement 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 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  $\{Y^1, \dots, Y^n\}$  from the common variance model (3), where  $K = (1, \dots, 1)^t$ ,  $\tau \sim \text{gamma}(1, 1)$ ,  $\phi$  is the identity function and  $\eta \in \mathbb{R}^d$  has independent components that are

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  $\eta$ , 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  $\mu_3, \mu_4$ , 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, 1999b), 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).

simulated from different univariate distributions that are summarized in Table 1. The draws for  $\eta_i$  are standardized such that  $\varepsilon_i = \tau\eta_i$  has mean zero and unit variance.

The matrix  $A$  is defined by  $A = R^T L$ , 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 whose 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  $W_n = \widehat{\Sigma}_n^{-1}$ .

As an alternative non-independent component method we include topographical ICA (TICA) of Hyvärinen, Hoyer and Inki (2001). 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, Hoyer and Inki, 2001, equation 3.10). The minimum distance estimators that we propose do not exploit this knowledge.

For comparison purposes we also include FastICA (Hyvärinen, 1999b), 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, Cichocki and Yang \(1996\)](#) error:

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

where  $a_{ij}$  is the  $i, j$  element of  $A_0 \hat{A}_{W_n}^{-1}$ . 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.

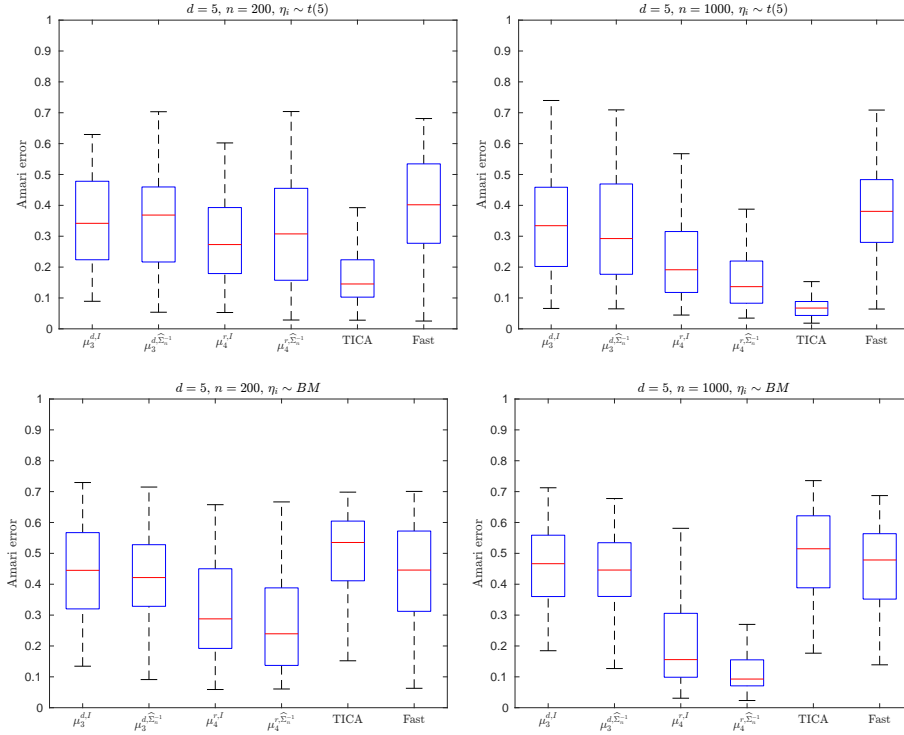
As an alternative non-independent component method, TICA works well when the true likelihood is close to the imposed objective function of TICA. Generally, this is the case for all densities that are similar to the Student's  $t$  density. When the true density is far from the approximating densities 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,  $\eta$  is drawn from the uniform distribution on the  $d$ -sphere and  $\tau$  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  $\tau$  and  $\eta$  are dependent and the diagonal tensor identification result no longer holds. That said, Corollary 5.15 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  $\eta$  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

Fig 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 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.

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  $\varepsilon$  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  $\varepsilon$ . 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.

Our results are practically important as they provide researchers with alternative identification and inference tools when independence fails to hold. Indeed, while the literature has produced a variety of tests for the independence of the components of  $\varepsilon$  in model (1) (e.g.

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  $\eta$ , 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  $\mu_3, \mu_4$ , 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, 1999b), 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).

Matteson and Tsay, 2017), it is unclear how one should proceed in practice when such tests reject. Our paper provides methods that can be applied in such situations.

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, Schennach and Zhang, 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

- AMARI, S.-I., CICHOCKI, A. and YANG, H. (1996). A New Learning Algorithm for Blind Signal Separation. *Advances in Neural Information Processing Systems* **8** 757–763.
- ASAI, M., MCALEER, M. and YU, J. (2006). Multivariate Stochastic Volatility: A Review. *Econometric Reviews* **25** 145–175.
- AZZALINI, A. and GENTON, M. G. (2008). Robust likelihood methods based on the skew-t and related distributions. *International Statistical Review* **76** 106–129.
- BACH, F. R. and JORDAN, M. I. (2002). Kernel Independent Component Analysis. *Journal of Machine Learning Research* **3** 1–48.
- BACH, F. R. and JORDAN, M. I. (2003). Beyond Independent Components: Trees and Clusters. *Journal of Machine Learning Research* **4** 1205–1233.

- BACK, A. D. and WEIGEND, A. S. (1997). A first application of independent component analysis to extracting structure from stock returns. *International Journal of Neural Systems* **8** 473–484.
- BEKAERT, G., ENGSTROM, E. and ERMOLOV, A. (2021). Macro risks and the term structure of interest rates. *Journal of Financial Economics* **141** 479–504.
- BEKAERT, G., ENGSTROM, E. and ERMOLOV, A. (2022). Identifying Aggregate Demand and Supply Shocks Using Sign Restrictions and Higher-Order Moments. working paper.
- CARDOSO, J.-F. (1989). Source separation using higher order moments. In *International Conference on Acoustics, Speech, and Signal Processing*, 2109–2112 vol.4.
- CARDOSO, J.-F. (1999). High-Order Contrasts for Independent Component Analysis. *Neural Computation* **11** 157–192.
- CARDOSO, J.-F. and SOULOUMIAC, A. (1993). Blind Beamforming for Non-Gaussian Signals. *IEEE Proceedings F - Radar and Signal Processing* **140**. <https://doi.org/10.1049/ip-f-2.1993.0054>
- CHAMBERLAIN, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* **34** 305–334.
- CHEN, A. and BICKEL, P. J. (2006). Efficient independent component analysis. *The Annals of Statistics* **34** 2825 – 2855.
- COMON, P. (1994). Independent component analysis, A new concept? *Signal Processing* **36**.
- COMON, P. and JUTTEN, C. (2010). *Handbook of Blind Source Separation*. Academic Press, Oxford.
- DARMOIS, G. (1953). Analyse générale des liaisons stochastiques: étude particulière de l'analyse factorielle linéaire. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute* **21** 2–8.
- DAVIS, R. and NG, S. (2022). Time Series Estimation of the Dynamic Effects of Disaster-Type Shocks Working paper.
- DRAUTZBURG, T. and WRIGHT, J. H. (2021). Refining Set-Identification in VARs through Independence. Working Paper No. 29316, National Bureau of Economic Research.
- ERICKSON, T., JIANG, C. H. and WHITED, T. M. (2014). Minimum distance estimation of the errors-in-variables model using linear cumulant equations. *Journal of Econometrics* **183** 211–221.
- ERIKSSON, J. and KOIVUNEN, V. (2003). Identifiability and Separability of Linear ICA models revisited. In *4th International Symposium on Independent Components Analysis and Blind Source Separation (ICA2003)* **1** 23–27.
- FORBES, F. and WRAITH, D. (2014). A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering. *Statistics and Computing* **24** 971–984.
- GEARY, R. C. (1941). Inherent Relations between Random Variables. *Proceedings of the Royal Irish Academy. Section A: Mathematical and Physical Sciences* **47** 63–76.
- GUAY, A. (2021). Identification of structural vector autoregressions through higher unconditional moments. *Journal of Econometrics* **225** 27–46.
- HALL, A. R. (2005). *Generalized Method of Moments*. Oxford University Press.
- HALLIN, M. and MEHTA, C. (2015). R-Estimation for Asymmetric Independent Component Analysis. *Journal of the American Statistical Association* **110** 218–232.
- HAN, F. and LIU, H. (2018). ECA: High-dimensional elliptical component analysis in non-Gaussian distributions. *Journal of the American Statistical Association* **113** 252–268.
- HANSEN, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica* **50** 1029–1054.
- HANSEN, L. P., HEATON, J. and YARON, A. (1996). Finite-Sample Properties of Some Alternative GMM Estimators. *Journal of Business & Economic Statistics* **14** 262–280.
- HANSEN, B. E. and LEE, S. (2021). Inference for Iterated GMM Under Misspecification. *Econometrica* **89** 1419–1447.
- HASTIE, T. and TIBSHIRANI, R. (2002). Independent Components Analysis through Product Density Estimation. In *Proceedings of the 15th International Conference on Neural Information Processing Systems. NIPS'02* 665–672. MIT Press, Cambridge, MA, USA.
- HAYASHI, F. (2000). *Econometrics*. Princeton University Press.
- HYVÄRINEN, A. (1999a). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks* **10** 626–634.
- HYVÄRINEN, A. (1999b). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks* **10** 626–634.
- HYVÄRINEN, A. (2013). Independent component analysis: recent advances. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **371** 20110534. <https://doi.org/10.1098/rsta.2011.0534>
- HYVÄRINEN, A., HOYER, P. O. and INKI, M. (2001). Topographic independent component analysis. *Neural computation* **13** 1527–1558.

- HYVÄRINEN, A., KARHUNEN, J. and OJA, E. (2001). *Independent Component Analysis*. Wiley, New York.
- ILMONEN, P. and PAINDAVEINE, D. (2011). Semiparametrically efficient inference based on signed ranks in symmetric independent component models. *The Annals of Statistics* **39** 2448 – 2476.
- ILMONEN, P., NORDHAUSEN, K., OJA, H. and OLLILA, E. (2010). A New Performance Index for ICA: Properties, Computation and Asymptotic Analysis. In *Latent Variable Analysis and Signal Separation* (V. VIGNERON, V. ZARZOSO, E. MOREAU, R. GRIBONVAL and E. VINCENT, eds.) 229–236. Springer Berlin Heidelberg, Berlin, Heidelberg.
- KELKER, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā: The Indian Journal of Statistics, Series A* 419–430.
- KILIAN, L. and LÜTKEPOHL, H. (2017). *Structural Vector Autoregressive Analysis*. Cambridge University Press.
- LANNE, M. and LUOTO, J. (2021). GMM Estimation of Non-Gaussian Structural Vector Autoregression. *Journal of Business & Economic Statistics* **39** 69–81.
- LEE, A. and MESTERS, G. (2021). Robust Non-Gaussian Identification and Inference for Simultaneous Equations. *Working Paper*.
- LEWBEL, A., SCHENNACH, S. M. and ZHANG, L. (2021). Identification of a Triangular Two Equation System Without Instruments. working paper.
- LIM, L.-H. (2021). Tensors in computations. *Acta Numerica* **30** 555–764.
- LUDVIGSON, S. C., MA, S. and NG, S. (2021). Uncertainty and Business Cycles: Exogenous Impulse or Endogenous Response? *American Economic Journal: Macroeconomics* **13**.
- LUKACS, E. (1958). Some extensions of a theorem of Marcinkiewicz. *Pacific Journal of Mathematics* **8** 487–501.
- MARCINKIEWICZ, J. (1939). Sur une propriété de la loi de Gauss. *Mathematische Zeitschrift* **44** 612–618.
- MARRON, J. S. and WAND, M. P. (1992). Exact mean integrated squared error. *The Annals of Statistics* 712–736.
- MATTESON, D. S. and TSAY, R. S. (2017). Independent Component Analysis via Distance Covariance. *Journal of the American Statistical Association* **112** 623–637.
- MCCULLAGH, P. (2018). *Tensor methods in statistics: Monographs on statistics and applied probability*. Chapman and Hall/CRC.
- MEYER-BASE, A., AUER, D. and WISMUELLER, A. (2003). Topographic independent component analysis for fMRI signal detection. In *Proceedings of the International Joint Conference on Neural Networks, 2003*. **1** 601–605. IEEE.
- MEYER-BÄSE, A., LANGE, O., WISMÜLLER, A. and RITTER, H. (2004). Model-free functional MRI analysis using topographic independent component analysis. *International journal of neural systems* **14** 217–228.
- MONTIEL OLEA, J. L., PLAGBORG-MØLLER, M. and QIAN, E. (2022). SVAR Identification from Higher Moments: Has the Simultaneous Causality Problem Been Solved? *AEA Papers and Proceedings* **112** 481–85.
- PALMER, J. A., KREUTZ-DELGADO, K., RAO, B. D. and MAKEIG, S. (2007). 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 97–104. Springer.
- ROSSELL, D. and ZWIERNIK, P. (2021). Dependence in elliptical partial correlation graphs. *Electronic Journal of Statistics* **15** 4236–4263.
- SAMWORTH, R. J. and YUAN, M. (2012). Independent component analysis via nonparametric maximum likelihood estimation. *The Annals of Statistics* **40** 2973 – 3002.
- SCHENNACH, S. M. (2021). Measurement systems. *Journal of Economic Literature*. forthcoming.
- SHIMIZU, S., HOYER, P. O., HYVÄRINEN, A. and KERMINEN, A. (2006). A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research* **7** 2003–2030.
- SKITOVIC, V. P. (1953). On a property of the normal distribution. *Dokl. Akad. Nauk SSSR (N.S.)* **89** 217–219.
- SPEED, T. P. (1983). Cumulants and partition lattices 1. *Australian Journal of Statistics* **25** 378–388.
- VELASCO, C. (2022). Identification and Estimation of Structural VARMA Models Using Higher Order Dynamics. *Journal of Business & Economic Statistics*. forthcoming.
- VOGEL, D. and FRIED, R. (2011). Elliptical graphical modelling. *Biometrika* **98** 935–951.
- ZWIERNIK, P. (2012). L-cumulants, L-cumulant embeddings and algebraic statistics. *Journal of Algebraic Statistics* **3** 11 – 43.
- ZWIERNIK, P. (2016). Semialgebraic statistics and latent tree models. *Monographs on Statistics and Applied Probability* **146** 146.