

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

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

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 21, 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 $\Omega = \{QA : Q \in O(d)\}$, 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 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 r th 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* 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 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.

¹For 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 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, 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

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

where τ is some positive random variable with 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

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

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 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_d^2 distribution, such that ε 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 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

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

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 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

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

where 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 $k = 2, 3, \dots$, 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 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 $k = 2, 3, \dots$ 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 $X = (X_1, \dots, X_d)'$ 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

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

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

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

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

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

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 $A^{(1)} \in \mathbb{R}^{m_1 \times n_1}, \dots, A^{(r)} \in \mathbb{R}^{m_r \times n_r}$:

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

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 $\text{var}(Y) = (A'A)^{-1}$ and so it is enough to narrow down potential candidates for A to the compact set

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

Our main insight is as follows: Since ε is unobserved, multiplying (1) by $Q \in \text{O}(d)$ 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 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

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

and for any given $Q \in \text{O}(d)$, $QA \in \Omega$ 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

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

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 $\text{var}(\varepsilon) = I_d$. 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 $\text{SP}(d)$. These are the $2^d d!$ matrices that are of the form DP , where $D, P \in \text{O}(d)$ 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 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 $T_{\mathbf{i}} = 0$ unless $\mathbf{i} = (i, \dots, i)$ for some $i = 1, \dots, d$.

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 $x = (x_1, \dots, x_d)$. 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. \square

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 $Q \in \text{SP}(d)$, i.e. $\mathcal{G}_T(\mathcal{V}) = \text{SP}(d)$.*

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

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

Since $r \geq 3$, this equality can hold only if $T_{j \dots j} = T_{k \dots k} = 0$, which is impossible by our genericity assumption. \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 $\frac{\partial^2}{\partial x_1^2} f_T(x) = \frac{\partial^2}{\partial x_2^2} f_T(x) = 0$. 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 $\mathbb{E}\varepsilon = 0$, $\text{var}(\varepsilon) = I_d$ 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 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 $\text{var}(\varepsilon) = I_d$. 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 $\phi : \mathbb{R} \rightarrow \mathbb{R}_{>0}$, 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, $h_r(\varepsilon)$ 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. \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.

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. \square

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

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

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

$$(13) \quad T_{+ \dots + ii} \neq T_{+ \dots + jj} \quad \text{for all } i \neq j.$$

Then $Q \bullet T$ is reflectionally invariant for $Q \in O(d)$ if and only if $Q \in SP(d)$, 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 $Q \in O(d)$ 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 $1 \leq i \leq j \leq d$

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

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 $r \geq 2$ 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

$$(14) \quad (Q, \dots, Q, I_d, I_d) \cdot T = (I_d, \dots, I_d, Q', Q') \cdot T.$$

We want to show that this equality implies that Q is a diagonal matrix. 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 (14) 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 $S = (S_{\mathbf{i}}) \in S^r(\mathbb{R}^d)$. Note however that S is a reflectionally invariant tensor. Indeed, if some index appears in \mathbf{i} odd number of times then $S_{\mathbf{i}} = T_{i_1 \dots i_r} = 0$ as the same index appears in (\mathbf{i}, u, u) odd number of times. Since T satisfies (13), S satisfies (13) too. Indeed,

$$\sum_{k_1} \cdots \sum_{k_{l-1}} S_{k_1 k_1 \dots k_{l-1} k_{l-1} i i} = \sum_{k_1} \cdots \sum_{k_{l-1}} \sum_{k_l} T_{k_1 k_1 \dots k_{l-1} k_{l-1} k_l k_l i i}$$

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. \square

Proof of Theorem 5.10. The left implication is clear. 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 5.12 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 $i = 1, \dots, d$, we conclude that $Q \in \text{SP}(d)$. \square

Combining Proposition 4.1 and Theorem 5.10 implies the following result.

Theorem 5.13. *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 (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 $\text{var}(\varepsilon) = I_d$. If ε follows the general common covariance model (3) then $h_4(\varepsilon)$ is reflectionally invariant. If ε follows 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

$$(15) \quad D_i(\tau \odot U) = \tau \odot (D_i U) \stackrel{d}{=} \tau \odot U,$$

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. \square

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 $i = 1, \dots, d$.

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

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

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 symmetry in the moments of $\eta \sim \mathcal{U}_d(\varepsilon)$ to show that $\mu_4(\varepsilon)$ is generic as long as

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

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 $\dim(\text{O}(d)) = \binom{d}{2}$, 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 5.16. If T is a generic tensor in $\mathcal{V}(\mathcal{I})$ then $Q \bullet T \in \mathcal{V}(\mathcal{I})$ if and only if $Q \in \text{SP}(d)$.

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 5.17. *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 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 $T_{\mathbf{i}}$ for $\mathbf{i} \in \mathcal{I}$, i.e. the set of zero restricted higher order tensor entries.

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

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

where 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 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 supplementary material Section S3 where we provide explicit computational formulas.

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

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

where W is an $d_g \times d_g$ positive definite weighting matrix, $\|v\|_W^2 = v'Wv$. The following result is 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 $Q \in \text{SP}(d)$.*

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 $A_0^{-1}A \in O(d)$, or in other words, $A = QA_0$ for some $Q \in O(d)$. 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 $\mathcal{G}_T(\mathcal{V}) = \text{SP}(d)$. \square

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

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

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

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

$$(19) \quad \Sigma = \lim_{n \rightarrow \infty} \text{var}(\sqrt{n}\hat{g}_n(QA_0))$$

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

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

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

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

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

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 $\hat{A}_{\hat{\Sigma}_n^{-1}}$. 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 $\hat{A}_{\hat{\Sigma}_n^{-1}}$ attains the semi-parametric efficiency bound in the class of non-parametric models characterized by restrictions $T_i = 0$ for $i \in \mathcal{I}$.

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 $\hat{A}_{\hat{\Sigma}_n^{-1}}$ 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 $W_n = \hat{\Sigma}_n(A)^{-1}$ as a weighting matrix, hence parametrizing the asymptotic variance estimate as a function of A , and minimize the objective function (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	$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
SKU	Skewed Unimodal	$\frac{1}{5}\mathcal{N}(0, 1) + \frac{1}{5}\mathcal{N}(\frac{1}{2}, (\frac{2}{3})^2) + \frac{3}{5}\mathcal{N}(\frac{13}{12}, (\frac{5}{9})^2)$
KU	Kurtotic Unimodal	$\frac{2}{3}\mathcal{N}(0, 1) + \frac{1}{3}\mathcal{N}(0, (\frac{1}{10})^2)$
BM	Bimodal	$\frac{1}{2}\mathcal{N}(-1, (\frac{2}{3})^2) + \frac{1}{2}\mathcal{N}(1, (\frac{2}{3})^2)$
SBM	Separated Bimodal	$\frac{1}{2}\mathcal{N}(-\frac{3}{2}, (\frac{1}{2})^2) + \frac{1}{2}\mathcal{N}(\frac{3}{2}, (\frac{1}{2})^2)$
SKB	Skewed Bimodal	$\frac{3}{4}\mathcal{N}(0, 1) + \frac{1}{4}\mathcal{N}(\frac{3}{2}, (\frac{1}{3})^2)$
TRI	Trimodal	$\frac{9}{20}\mathcal{N}(-\frac{6}{5}, (\frac{3}{5})^2) + \frac{9}{20}\mathcal{N}(\frac{6}{5}, (\frac{3}{5})^2) + \frac{1}{10}\mathcal{N}(0, (\frac{1}{4})^2)$
CL	Claw	$\frac{1}{2}\mathcal{N}(0, 1) + \sum_{l=0}^4 \frac{1}{10}\mathcal{N}(l/2 - 1, (\frac{1}{10})^2)$
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.

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 $K = (1, \dots, 1)'$, $\tau \sim \text{gamma}(1, 1)$, 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 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 $W_n = \hat{\Sigma}_n^{-1}$.

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:

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

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.

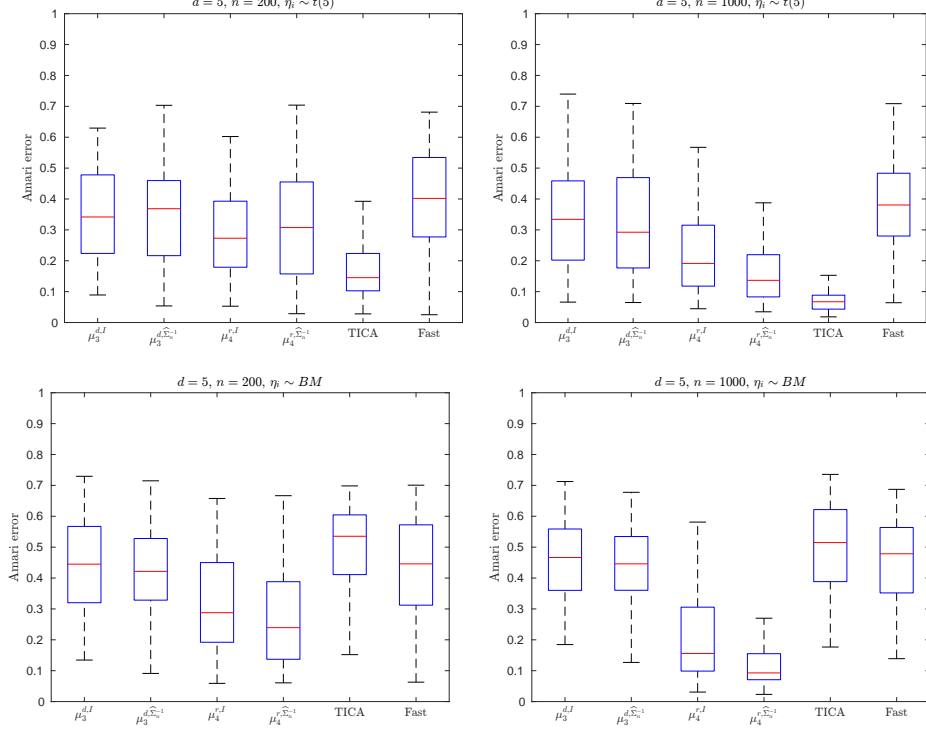
Minimum distance methods based on third order diagonal moment restrictions perform well when the true density has strong skewness (e.g. SKU). When used with the efficient weighting matrix this minimum distance method nearly always outperform the ICA methods, but even then the reflectionally invariant 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, 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 [Schen-nach, 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 Darmais. 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 Erkki 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. doi: 10.1007/978-3-642-15995-4_29.
- 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, Ddrotsee 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`

SUPPLEMENTARY MATERIAL: NON-INDEPENDENT COMPONENTS ANALYSIS

GEERT MESTERS AND PIOTR ZWIERNIK

ABSTRACT. We provide the following additional results.

- S1:** Additional motivation
- S2:** Local identification beyond signed-permutations
- S3:** Moments and Cumulants — some useful properties
- S4:** Additional inference tools
- S5:** Computing the asymptotic variance
- S6:** Additional simulation results
- S7:** Omitted proofs

S1. ADDITIONAL MOTIVATION

In this section we discuss some additional relations that aim to further highlight the usefulness of the identification results presented in the main text.

S1.1. Scaled Elliptical LiNGAM. For the model $AY = \varepsilon$, where the elements of ε are independent and non-Gaussian [Shimizu et al. \[2006\]](#), showed that one can uniquely recover A if there exists an (unknown) permutation of the rows of A that is lower triangular, i.e. the model corresponds to a directed acyclic graph. The proposed LiNGAM discovery algorithm uses ICA and a search over permutations to find the best fitting lower triangular model.

Now reconsider the multiple scaled elliptical components model

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

with $\tau \in \mathbb{R}^d$ and U independent.

With elliptical errors the LiNGAM algorithm can no longer be used. However, the results of this paper suggests a natural modification where the ICA algorithm is replaced by the moment or cumulant based estimation methods that we introduce in Section 6. These methods are build on the new identification results for the multiple scaled elliptical components model.

Specifically, in Algorithm A of [Shimizu et al. \[2006\]](#) one can replace the ICA method that is used in step 1 by the minimum distance moment/cumulant estimation method of Section 6. The other steps of the algorithm do not require adjustment.

Date: September 21, 2023.

2020 Mathematics Subject Classifications. 15A69, 62H99.

S1.2. Invariance. In Section 2 we motivated non-independent component models using specific examples (e.g. common variance model) as well as by relaxing independence (e.g. mean independence). In both cases the resulting model still implied sufficient zero restrictions on the higher order moments/cumulants of ε to ensure the identifiability of A (cf Corollaries 5.7 and 5.14). Here we briefly show that such zero restrictions can also arise from invariance properties of the distribution of ε .

Suppose that the distribution of ε is the same as the distribution of $D\varepsilon$ for every diagonal matrix D with $D_{ii} = \pm 1$ for all $i = 1, \dots, d$ (e.g. when ε has spherical distribution). In this case, by multilinearity of cumulants,

$$[h_r(D\varepsilon)]_{i_1 \dots i_r} = D_{i_1 i_1} \cdots D_{i_r i_r} [h_r(\varepsilon)]_{i_1 \dots i_r}.$$

Since D is arbitrary, $[\kappa_r(\varepsilon)]_{i_1 \dots i_r}$ must be zero unless all indices appear even number of times. In particular, if r must be even and for example, if $r = 4$, the only potentially non-zero cumulants are κ_{iiii} and κ_{iijj} . These zero patterns correspond exactly with the reflectionally invariant restrictions introduced in Section 5.2 and as such Corollary 5.14 also ensure the identifiability of A in $AY = \varepsilon$ when the distribution of ε is the same as the distribution of $D\varepsilon$.

S1.3. Alternative estimation methods. In the main text we outlined some minimum distance estimation methods for estimating A in $AY = \varepsilon$ based on the identifying moment/cumulant restrictions. We adopted this approach as it can be implemented naturally based on our identification results. That said, for specific non-independent components models it is obviously feasible to develop alternative estimators based on the identification results. To illustrate, we discuss some approaches for the mean independent components model:

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

Shao and Zhang [2014] introduce *martingale difference correlations* to measure the departure of conditional mean independence between a scalar response variable (i.e. ε_i) and a vector predictor variable (i.e. ε_{-i}). This metric is a natural extension of distance correlation proposed by Székely et al. [2007], which was adopted in Matteson and Tsay [2017] for independent components analysis. These observations immediately suggest that jointly minimizing the martingale difference correlations between ε_i and ε_{-i} for all i with respect to A provides an attractive approach for estimating mean independent components models.

Alternatively, recall that the efficient ICA method of Chen and Bickel [2006] is based on setting the efficient score function of the semi-parametric ICA model (with independent errors) to zero. The analytical form of these efficient scores relies on the independence assumption. When relaxing towards mean independence it is straightforward to derive a new analytical expression for the efficient scores and apply the algorithm of Chen and Bickel [2006] to set these scores to zero.

S2. LOCAL IDENTIFICATION BEYOND SIGNED-PERMUTATIONS

The results in Section 5 stipulate conditions on moment tensors $T = \mu_r(\varepsilon)$ or cumulant tensors $T = \kappa_r(\varepsilon)$ for which A can be recovered up to sign and permutation. This section gives minimal conditions on \mathcal{V} that ensure that \mathcal{G}_T is finite. We subsequently use this result to highlight the gap that exists between restrictions that lead to finite sets and restrictions that lead to signed permutation sets. This finding has the 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.

Let $\mathcal{V} \subset S^r(\mathbb{R}^d)$ be a set given as a set of zeros of a system of polynomials in the coordinates of $S^r(\mathbb{R}^d)$ (such set is called an algebraic variety). A subset $\mathcal{U} \subseteq \mathcal{V}$ is Zariski open in \mathcal{V} if the complement $\mathcal{V} \setminus \mathcal{U}$ is also an algebraic variety. In particular, a Zariski open set is also open in the classical topology. For example, the set of diagonal tensors in $S^r(\mathbb{R}^d)$ with at most one zero on the diagonal forms a Zariski open subset of the set of diagonal tensors. Similarly, the set of reflectionally invariant tensors satisfying the genericity condition (13) is Zariski open in the set of reflectionally invariant tensors. Note that, in both cases, the constraints defining \mathcal{V} and $\mathcal{V} \setminus \mathcal{U}$ were linear.

Recall from (10) that for $T = h_r(\varepsilon) \in \mathcal{U}$ we define $\mathcal{G}_T(\mathcal{U}) = \{Q \in O(d) : Q \bullet T \in \mathcal{U}\}$ to be the set of all orthogonal matrices for which $h_r(Q\varepsilon)$ also lies in \mathcal{U} .

Definition S1. The problem of recovering A in (1) is locally identifiable under moment/cumulant constraints $\mathcal{U} \subseteq \mathcal{V} \subset S^r(\mathbb{R}^d)$ with \mathcal{U} open in \mathcal{V} if every point of $\mathcal{G}_T(\mathcal{U})$ is an isolated point of $\mathcal{G}_T(\mathcal{U})$.

Note that, at least in principle, $\mathcal{G}_T(\mathcal{U})$ could contain infinitely many isolated points. The following result establishes link between local identification and finiteness of \mathcal{G}_T .

Proposition S2. Let \mathcal{U} be a Zariski open subset of \mathcal{V} . For $T^* \in \mathcal{U}$ we have $|\mathcal{G}_{T^*}(\mathcal{U})| < \infty$ if and only if each point of $\mathcal{G}_{T^*}(\mathcal{U})$ is an isolated point of $\mathcal{G}_{T^*}(\mathcal{U})$.

Proof. The right implication is clear. For the left implication first note that $\mathcal{G}_{T^*}(\mathcal{U})$ is a Zariski open subset of the real algebraic variety $\mathcal{G}_{T^*}(\mathcal{V})$. Indeed, if $f_1(T) = \dots = f_k(T) = 0$ are the polynomials, in T , describing \mathcal{V} then the polynomials, in Q , describing $\mathcal{G}_{T^*}(\mathcal{V})$ within $O(d)$ are $f_1(Q \bullet T^*) = \dots = f_k(Q \bullet T^*) = 0$. Similarly, if $\mathcal{V} \setminus \mathcal{U}$ is described within \mathcal{V} by $g_1(T) = \dots = g_l(T) = 0$. Then $\mathcal{G}_{T^*}(\mathcal{V}) \setminus \mathcal{G}_{T^*}(\mathcal{U})$ is described by $g_1(Q \bullet T^*) = \dots = g_l(Q \bullet T^*) = 0$.

Since $\mathcal{G}_{T^*}(\mathcal{V})$ is a real algebraic variety, the set of its isolated points is equal to its zero-dimensional components and so it must be finite; see for example Theorem 4.6.2 in Cox et al. [2013]. It is then enough to show that if Q° is isolated in $\mathcal{G}_{T^*}(\mathcal{U})$ then it must be isolated in $\mathcal{G}_{T^*}(\mathcal{V})$. Suppose that $Q^\circ \in \mathcal{G}_{T^*}(\mathcal{U})$ is not isolated in $\mathcal{G}_{T^*}(\mathcal{V})$. Then it must lie on an irreducible

component of the variety $\mathcal{G}_{T^*}(\mathcal{V})$ of a positive dimension. By assumption, for this Q° , $g_1(Q^\circ \bullet T^*) \neq 0, \dots, g_l(Q^\circ \bullet T^*) \neq 0$. Thus, in any sufficiently small neighbourhood of Q° there will be a point that lies in $\mathcal{G}_{T^*}(\mathcal{V})$ and g_1, \dots, g_l evaluate to something non-zero. In other words, in any sufficiently small neighbourhood of Q° there is a point in $\mathcal{G}_{T^*}(\mathcal{U})$ proving that Q° cannot be isolated in $\mathcal{G}_{T^*}(\mathcal{U})$, which leads to contradiction. \square

Remark S3. The proof of Proposition S2 also shows that if \mathcal{U} is a Zariski open subset of \mathcal{V} and $T \in \mathcal{U}$ then $\mathcal{G}_T(\mathcal{U})$ is a Zariski open subset of $\mathcal{G}_T(\mathcal{V})$. Moreover, $Q \in \mathcal{G}_T(\mathcal{U})$ is isolated if and only if it is isolated in $\mathcal{G}_T(\mathcal{V})$.

By the above remark, to show local identifiability it is enough to show that every element of $\mathcal{G}_T(\mathcal{U})$ is isolated in $\mathcal{G}_T(\mathcal{V})$. To show this, we take any point in $\mathcal{G}_T(\mathcal{U})$ and try to perturb it infinitesimally staying in the orthogonal group. We need that every such infinitesimal perturbation sends the point outside of $\mathcal{G}_T(\mathcal{V})$.

Lemma S4. *For a fixed $T \in S^r(\mathbb{R}^d)$, consider the map from $\mathbb{R}^{d \times d}$ to $S^r(\mathbb{R}^d)$ given by $A \mapsto A \bullet T$. Its derivative at A is a linear mapping on $\mathbb{R}^{d \times d}$ defined by*

$$(S1) \quad K_{T,A}(V) = (V, A, \dots, A) \bullet T + \dots + (A, \dots, A, V) \bullet T.$$

Moreover, if A is invertible, then

$$(S2) \quad K_{T,A}(V) = K_{A \bullet T, I_d}(VA^{-1}).$$

Proof. For any direction $V \in \mathbb{R}^{d \times d}$, we have

$$\begin{aligned} (A + tV) \bullet T &= A \bullet T \\ &= t(V, A, \dots, A) \bullet T + \dots + t(A, \dots, A, V) \bullet T + o(t). \end{aligned}$$

So the proof of the first claim follows by the definition of a derivative. The second claim follows by direct calculation. \square

For a given linear subspace $\mathcal{V} \subseteq S^r(\mathbb{R}^d)$, let $\pi_{\mathcal{V}} : S^r(\mathbb{R}^d) \rightarrow \mathcal{V}^\perp$ denote the orthogonal projection on \mathcal{V}^\perp . Of course, $T \in \mathcal{V}$ if and only if $\pi_{\mathcal{V}}(T) = 0$. Moreover, if $\mathcal{V} = \mathcal{V}(\mathcal{I})$ is given by zero constraints, then $\pi_{\mathcal{V}}(T)$ simply gives the coordinates T_i for $i \in \mathcal{I}$.

In the next result, $K_{I_d, A}(V) = (V, A) \bullet I_d + (A, V) \bullet I_d$, which is a special instance of (S1) for $r = 2$.

Lemma S5. *Let \mathcal{U} be a Zariski open subset of \mathcal{V} . A point Q is an isolated point of $\mathcal{G}_T(\mathcal{U})$ if and only if*

$$(S3) \quad K_{I_d, Q}(V) = 0 \text{ and } \pi_{\mathcal{V}}(K_{T, Q}(V)) = 0 \quad \text{implies } V = 0.$$

Proof. Since,

$$(Q + tV)(Q + tV)' = I_d + t(VQ' + QV') + o(t),$$

V is a direction in the tangent space to $O(d)$ at Q if and only if $VQ' + QV' = 0$. Equivalently,

$$VQ' + QV' = (V, Q) \bullet I_d + (Q, V) \bullet I_d = K_{I_d, Q}(V) = 0.$$

Thus, the first condition $K_{I_d, Q}(V) = 0$ simply restates that V lies in the tangent space of $O(d)$ at Q .

The proof of Proposition S2 showed that, $\mathcal{U} \subseteq \mathcal{V}$ is Zariski open, then $\mathcal{G}_T(\mathcal{U})$ is Zariski open (and so also open in the classical topology) in $\mathcal{G}_T(\mathcal{U})$. Thus, if Q is not isolated, every neighborhood of Q must contain an element in $\mathcal{G}_T(\mathcal{U})$ different than Q . In other words, the point $Q \in \mathcal{G}_T(\mathcal{U})$ is not isolated if and only if there exists a tangent direction $V \neq 0$ such that

$$\pi_{\mathcal{V}}((Q + tV) \bullet T) - \pi_{\mathcal{V}}(Q \bullet T) = \pi_{\mathcal{V}}((Q + tV) \bullet T) = o(t).$$

Taking the limit $t \rightarrow 0$, we get that equivalently $\pi_{\mathcal{V}}(K_{T, Q}(V)) = 0$. This shows that Q is isolated if and only if no such non-trivial tangent direction exists. \square

Remark S6. In the examples of Section 5, for $T \in \mathcal{U} \subseteq \mathcal{V}$, we always had $\mathcal{G}_T(\mathcal{U}) = \mathcal{G}_T(\mathcal{V}) = \text{SP}(d)$. The proof of Proposition S2 suggests that, at least in principle $\mathcal{G}_T(\mathcal{U})$ could be finite but $\mathcal{G}_T(\mathcal{V})$ could have components of positive dimension. In the proof of the next result, we crucially rely on the fact that we compute $\mathcal{G}_T(\mathcal{U})$ rather than $\mathcal{G}_T(\mathcal{V})$.

Note that the dimension of the orthogonal group $O(d)$ is $\binom{d}{2}$, which is then also the minimal number of constraints that need to be imposed in order to hope for identifiability. The main result of this section studies local identifiability with a model defined by the minimal number of $\binom{d}{2}$ constraints with

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

We write $\mathcal{V}^\circ = \mathcal{V}(\mathcal{I})$. Denote

$$(S4) \quad B^{(j)} = [T_{klj \dots j}]_{k, l < j} \in S^2(\mathbb{R}^{j-1})$$

and define $\mathcal{U}^\circ \subset S^r(\mathbb{R}^d)$ as the set of tensors $T \in \mathcal{V}^\circ$ such that,

$$(S5) \quad \det \left(T_{j \dots j} I_{j-1} - (r-1) B^{(j)} \right) \neq 0 \quad \text{for all } j = 2, \dots, d.$$

Theorem S7. *If $T \in \mathcal{U}^\circ$ then $|\mathcal{G}_T(\mathcal{U}^\circ)| < \infty$.*

Proof. By Proposition S2 it is enough to show that each point of $\mathcal{G}_T(\mathcal{U}^\circ)$ is isolated. By Lemma S5, equivalently for every $Q \in \mathcal{G}_T(\mathcal{U}^\circ)$, if $K_{I_d, Q}(V) = 0$ and $\pi_{\mathcal{V}}(K_{T, Q}(V)) = 0$ then $V = 0$. By (S2), $K_{I_d, Q}(V) = K_{I_d, I_d}(VQ')$. Thus, denoting $U = VQ'$, this condition is equivalent to saying that U antisymmetric ($U + U' = 0$). We will show that the conditions above imply that U must be zero. By assumption, we have $U_{ii} = 0$ and $U_{ij} = -U_{ji}$ for all $i \neq j$. Again using (S2), we get $\pi_{\mathcal{V}}(K_{T, Q}(V)) = \pi_{\mathcal{V}}(K_{Q \bullet T, I_d}(U))$. Denote $S := Q \bullet T$. Since $Q \in \mathcal{G}_T(\mathcal{U}^\circ)$, in particular, $S \in \mathcal{U}^\circ$. The condition $\pi_{\mathcal{V}}(K_{S, I_d}(U)) = 0$ means that for every $\mathbf{i} = (i, j, \dots, j)$ with $i < j$,

$(K_{S,I_d}(U))_{ij\dots j} = 0$. More explicitly,

$$\begin{aligned}
0 &= \sum_{l=1}^d U_{il} S_{lj\dots j} + \sum_{l=1}^d U_{jl} S_{ilj\dots j} + \dots + \sum_{l=1}^d U_{jl} S_{ij\dots jl} \\
&= U_{ij} S_{j\dots j} + (r-1) \sum_{l=1}^d U_{jl} S_{ilj\dots j} \\
&= -U_{ji} S_{j\dots j} + (r-1) \sum_{l=1}^d U_{jl} S_{ilj\dots j}
\end{aligned}$$

Let $u_j = (U_{j1}, \dots, U_{jj-1})$ for $j = 2, \dots, d$. Let first $j = d$. Using the matrix $B^{(d)}$ defined in (S4) the equation above gives

$$(S_{d\dots d} I_{d-1} - (r-1) B^{(d)}) u_d = 0.$$

This has a unique solution $u_d = 0$ if and only if $\det(S_{d\dots d} I_{d-1} - (r-1) B^{(d)}) \neq 0$, which holds by (S5). We have shown that the last row of U is zero. Now suppose that we have established that the rows $j+1, \dots, d$ of U are zero. If $j = 1$, we are done by the fact that U is antisymmetric. So assume $j \geq 2$. We will use the fact that $U_{jl} = 0$ if $l \geq j$. For every $i < j$

$$0 = -U_{ji} S_{j\dots j} + (r-1) \sum_{l \neq j} U_{jl} B_{il}^{(j)} = -U_{ji} S_{j\dots j} + (r-1) \sum_{l < j} B_{il}^{(j)} U_{lj}.$$

This again has a unique solution if and only if $\det(S_{j\dots j} I_{j-1} - (r-1) B^{(j)}) \neq 0$, which holds by (S5). Using a recursive argument, we conclude that $U = 0$. \square

Example S8. Consider $\mathcal{V}^\circ \subseteq S^3(\mathbb{R}^2)$ given by $T_{122} = 0$. Direct calculations show that, for any given generic T , there are 12 orthogonal matrices such that $Q \bullet T \in \mathcal{V}$. There are four elements given by the diagonal matrices together with 8 additional elements that depend on T . So, for example, if $T_{111} = 1$, $T_{222} = 2$, and $T_{112} = 3$ then the twelve elements are the four matrices D and eight matrices of the form

$$\frac{1}{5} D \begin{bmatrix} 3 & 4 \\ 4 & -3 \end{bmatrix} \quad \text{and} \quad \frac{1}{\sqrt{2}} D \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Going back to our original motivation, suppose ε is a two-dimensional mean-zero random vector with $\text{var}(\varepsilon) = I_2$. If we impose in addition that $\mathbb{E} \varepsilon_1 \varepsilon_2^2 = 0$, then, even if we impose some genericity conditions, the matrix A in (1) is identified only up to the set of 12 elements. Moreover, as illustrated above, these elements may look nothing like A in the sense that they are not obtained by simple row permutation and sign swapping.

Remark S9. The set $\mathcal{G}_T(\mathcal{U}^\circ)$ is finite but, as illustrated by Example S8, it typically contains matrices that do not have an easy interpretation. In particular, if $d = 2$ then \mathcal{V}° is given by a single constraint $T_{12\dots 2} = 0$. In

this case we can show that there are generically $4r$ *complex* solutions (which generalized the number 12 in the above example). There are 4 solutions given by the elements of \mathbb{Z}_2^2 and $4(r-1)$ extra solutions, which do not have any particular structure.

We conclude the following result.

Theorem S10. *Consider the model (1) with $\mathbb{E}\varepsilon = 0$, $\text{var}(\varepsilon) = I_d$ and suppose that either $\mu_r(\varepsilon) \in \mathcal{U}^\circ$ or $\kappa_r(\varepsilon) \in \mathcal{U}^\circ$. Then A is locally identifiable.*

S3. MOMENTS AND CUMULANTS — SOME USEFUL PROPERTIES

We collect some results on moments and cumulants and their sample estimates that are used below for some of the proofs.

S3.1. Combinatorial relationship between moments and cumulants.

Let Π_r be the poset of all set partitions of $\{1, \dots, r\}$ ordered by refinement. For $\pi \in \Pi_r$ we write $B \in \pi$ for a block in π . The number of blocks of π is denoted by $|\pi|$. For example, if $r = 3$ then Π_3 has 5 elements: 123, 1/23, 2/13, 3/12, 1/2/3. They have 1, 2, 2, 2, and 3 blocks respectively. If $\mathbf{i} = (i_1, \dots, i_r)$ then \mathbf{i}_B is a subvector of \mathbf{i} with indices corresponding to the block $B \subseteq \{1, \dots, r\}$. For any multiset $\{i_1, \dots, i_r\}$ of the indices $\{1, \dots, d\}$ we can relate the moments $\mu_r(Y)$ to the cumulants [e.g. [Speed, 1983](#)].

$$(S6) \quad [\mu_r(Y)]_{i_1, \dots, i_r} = \sum_{\pi \in \Pi_r} \prod_{B \in \pi} [\kappa_{|B|}(Y)]_{\mathbf{i}_B} ,$$

where B loops over each block in a given partition π . For instance, for $r = 3$ we have

$$[\mu_r(Y)]_{i_1, i_2, i_3} = \kappa_{i_1 i_2 i_3} + \kappa_{i_1 i_2} \kappa_{i_3} + \kappa_{i_1 i_3} \kappa_{i_2} + \kappa_{i_2 i_3} \kappa_{i_1} + \kappa_{i_1} \kappa_{i_2} \kappa_{i_3} ,$$

where we use the more convenient notation $\kappa_{i_1 \dots i_l} = [\kappa_l(Y)]_{i_1 \dots i_l}$. Similarly, from [Speed \[1983\]](#) we have

$$(S7) \quad [\kappa_r(Y)]_{i_1, \dots, i_r} = \sum_{\pi \in \Pi_r} (-1)^{|\pi|-1} (|\pi|-1)! \prod_{B \in \pi} [\mu_{|B|}(Y)]_{\mathbf{i}_B} .$$

For example,

$$[\kappa_r(Y)]_{i_1, i_2, i_3} = \mu_{i_1 i_2 i_3} - \mu_{i_1} \mu_{i_2 i_3} - \mu_{i_2} \mu_{i_1 i_3} - \mu_{i_3} \mu_{i_1 i_2} + 2\mu_{i_1} \mu_{i_2} \mu_{i_3} ,$$

using $\mu_{i_1 \dots i_l} = [\mu_l(Y)]_{i_1 \dots i_l}$.

The coefficients $(-1)^{|\pi|-1} (|\pi|-1)!$ in (S7) have an important combinatorial interpretation, which we now briefly explain. If \mathbf{P} is a finite partially ordered set (poset) with ordering \leq we define the zeta function on $\mathbf{P} \times \mathbf{P}$ as $\zeta(x, y) = 1$ if $x \leq y$ and $\zeta(x, y) = 0$ otherwise. The Möbius function is then defined by setting $\mathbf{m}(x, y) = 0$ if $x \not\leq y$ and

$$\sum_{x \leq z \leq y} \mathbf{m}(x, z) \zeta(z, y) = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{otherwise.} \end{cases}$$

Fixing a total ordering on \mathbf{P} , we can represent the zeta function by a matrix Z and then the matrix M representing the Möbius function is simply the inverse of Z . If this total ordering is consistent with the partial ordering of \mathbf{P} then both Z and M are upper-triangular and have ones on the diagonal; see Section 4.1 in [Zwiernik \[2016\]](#) for more details.

For the poset $\mathbf{\Pi}_r$ the Möbius function satisfies for any $\rho \leq \pi$ (ρ is a refinement of π)

$$(S8) \quad \mathbf{m}(\rho, \pi) = (-1)^{|\rho| - |\pi|} \prod_{B \in \pi} (|\rho_B| - 1)!,$$

where $|\rho_B|$ is the number of blocks in which ρ subdivides the block B of π . In particular, denoting by $\mathbf{1} \in \mathbf{\Pi}_r$ the one-block partition, for every $\pi \in \mathbf{\Pi}_r$

$$\mathbf{m}(\pi, \mathbf{1}) = (-1)^{|\pi| - 1} (|\pi| - 1)!.$$

To explain how $\mathbf{m}(\pi, \mathbf{1})$ appears in (S7), we recall the Möbius inversion formula, which becomes clear given the matrix formulation using Z and $M = Z^{-1}$.

Lemma S1 (Möbius inversion theorem). *Let \mathbf{P} be a poset. For two functions c, d on \mathbf{P} , we have $d(x) = \sum_{y \leq x} c(y)$ for all $x \in \mathbf{P}$ if and only if $c(x) = \sum_{y \leq x} \mathbf{m}(x, y) d(y)$.*

For example, this result gives the simple formula (S6) that defines moments in terms of cumulants.

S3.2. Laws of total expectation and cumulance. The law of total expectation is well known; for two random variables X, H defined on the same probability space we have $\mathbb{E}X = \mathbb{E}[\mathbb{E}(X|H)]$. [Brillinger \[1969\]](#) derives an analog result for cumulants.

Proposition S2 (Multivariate law of total cumulants). *Let $\kappa_s(X|H)$ be the conditional s -th cumulant tensor of X given a variable H . We have*

$$\kappa_r(X) = \sum_{\pi \in \mathbf{\Pi}_r} \text{cum}((\kappa_{|B|}(X|H))_{B \in \pi}),$$

where for $\mathbf{i} = (i_1, \dots, i_r)$

$$[\text{cum}((\kappa_{|B|}(X|H))_{B \in \pi})]_{\mathbf{i}} = \text{cum}((\text{cum}(X_{\mathbf{i}_B}|H))_{B \in \pi}).$$

It is certainly hard to parse this formula at first so we offer a short discussion. The expression $\text{cum}((\text{cum}(X_{\mathbf{i}_B}|H))_{B \in \pi})$ on the right denotes the cumulant of order $|\pi|$ of the conditional variances $\text{cum}(X_{\mathbf{i}_B}|H)$ for $B \in \pi$. A special case of this result is the law of total covariance.

$$[\kappa_2(X)]_{ij} = \text{cov}(X_i, X_j) = \mathbb{E}(\text{cov}(X_i, X_j|H)) + \text{cov}(\mathbb{E}(X_i|H), \mathbb{E}(X_j|H)),$$

where the first summand on the right corresponds to the partition 12 and the second corresponds to the split 1/2. Since there are five possible partitions

of $\{1, 2, 3\}$ the third order cumulant can be given in conditional cumulants as

$$\begin{aligned} [\kappa_3(X)]_{ijk} &= \mathbb{E}(\text{cum}(X_i, X_j, X_k|H)) + \text{cov}(\mathbb{E}(X_i|H), \text{cov}(X_j, X_k|H)) \\ &+ \text{cov}(\mathbb{E}(X_j|H), \text{cov}(X_i, X_k|H)) + \text{cov}(\mathbb{E}(X_k|H), \text{cov}(X_i, X_j|H)) \\ &+ \text{cum}(\mathbb{E}(X_i|H), \mathbb{E}(X_j|H), \mathbb{E}(X_k|H)). \end{aligned}$$

Proposition [S2](#) is useful for example if the components of X are conditionally independent given H in which case all mixed conditional cumulants vanish. Another scenario is when X conditionally on H is Gaussian, in which case all higher order conditional tensors vanish.

S3.3. Estimating moments and cumulants. Given a sample $\{Y_s\}_{s=1}^n$, unbiased estimates for the r th order moment tensor $\mu_r(Y)$ are obtained by computing the sample moments

$$(S9) \quad [\hat{\mu}_r]_{i_1 \dots i_r} = \frac{1}{n} \sum_{s=1}^n Y_{s,i_1} Y_{s,i_2} \dots Y_{s,i_r}.$$

Using our multilinear notation we can more compactly write

$$(S10) \quad \hat{\mu}_r = \frac{1}{n} \mathbf{Y}' \bullet I_r \in S^r(\mathbb{R}^d).$$

where $I_r \in S^r(\mathbb{R}^n)$ is the identity tensor, that is, the diagonal tensor satisfying $(I_r)_{t \dots t} = 1$ for all $1 \leq t \leq n$.

Unbiased estimates for the cumulants are computed using multivariate k-statistics [Speed \[1983\]](#), which generalize classical k-statistics introduced by [Fisher \[1930\]](#). For a collection of useful results on k-statistics see also [McCullagh, 2018](#), Chapter 4].

Specifically, the entries of the r th order k-statistic used to estimate the cumulant $[\kappa_r(Y)]_{i_1 \dots i_r}$ are given by (see [McCullagh, 2018](#), (4.5)-(4.7))

$$(S11) \quad [k_r]_{i_1, \dots, i_r} = \frac{1}{n} \sum_{t_1=1}^n \dots \sum_{t_r=1}^n \Phi_{t_1, \dots, t_r} Y_{t_1, i_1} \dots Y_{t_r, i_r}$$

with $\Phi \in S^r(\mathbb{R}^n)$ satisfying

$$\Phi_{t_1 \dots t_r} = (-1)^{\nu-1} \frac{1}{\binom{n-1}{\nu-1}},$$

where $\nu \leq n$ is the number of distinct indices in (t_1, \dots, t_r) . Let $\mathbf{Y} \in \mathbb{R}^{n \times d}$ be the data matrix. More compactly, we have

$$(S12) \quad k_r = \frac{1}{n} \mathbf{Y}' \bullet \Phi \in S^r(\mathbb{R}^d).$$

We note the following important result; see Proposition 4.3 in [Speed \[1986\]](#).

Proposition S3. *The k-statistic in (S11) forms a U-statistic. In particular, it is unbiased and it has the minimal variance among all unbiased estimators.*

Besides being unbiased and efficient, an additional benefit of working with \mathbf{k}_r statistics is that there are several statistical packages available that compute them, e.g. `kStatistics` for R and `PyMoments` for Python. The first package uses the powerful machinery of umbral calculus to make the symbolic computations efficient [Di Nardo et al. \[2009\]](#).

S3.4. \mathbf{k} -statistics and sample cumulants. For later considerations we need to understand better the relation between \mathbf{k}_r and the natural plug-in estimator $\hat{\mathbf{k}}_r$, which is obtained by first estimating the raw moments and then plugging them into (S7). The relevant sample moments that allow to compute $\hat{\mathbf{k}}_r$ from (S7) are summarized in $\hat{\boldsymbol{\mu}}_p$ for $p \leq r$.

If $B \subseteq [n]$ then write I_B for the identity tensor in $S^{|B|}(\mathbb{R}^n)$. For any partition $\pi \in \mathbf{\Pi}_r$ the tensor product $\bigotimes_{B \in \pi} I_B \in S^r(\mathbb{R}^n)$ satisfies

$$\left[\bigotimes_{B \in \pi} I_B \right]_{t_1 \dots t_r} = \prod_{B \in \pi} [I_B]_{t_B} = \begin{cases} 1 & t_i = t_j \text{ whenever } i, j \in B \in \pi, \\ 0 & \text{otherwise.} \end{cases}$$

For every $\pi \in \mathbf{\Pi}_r$, define coefficients

$$(S13) \quad c(\pi) = \sum_{\rho \leq \pi} \mathbf{m}(\rho, \pi) (-1)^{|\rho|-1} \frac{1}{\binom{n-1}{|\rho|-1}} = n \sum_{\rho \leq \pi} \mathbf{m}(\rho, \pi) \mathbf{m}(\rho, \mathbf{1}) \frac{1}{(n)_{|\rho|}},$$

where \mathbf{m} is the Möbius function on $\mathbf{\Pi}_r$ given in (S8) and $(n)_k = n(n-1) \cdots (n-k+1)$ is the corresponding falling factor.

Lemma S4. *We have*

$$\Phi = \sum_{\pi \in \mathbf{\Pi}_r} c(\pi) \bigotimes_{B \in \pi} I_B,$$

which gives an alternative formula for k -statistics

$$[\mathbf{k}_r]_{i_1, \dots, i_r} = \sum_{\pi \in \mathbf{\Pi}_r} n^{|\pi|-1} c(\pi) \prod_{B \in \pi} \hat{\boldsymbol{\mu}}_{i_B}.$$

Proof. For any t_1, \dots, t_r let ν be the number of distinct elements in this sequence and let π^* be the partition $[r]$ with ν blocks corresponding to indices that are equal. We have

$$\left(\sum_{\pi \in \mathbf{\Pi}_r} c(\pi) \bigotimes_{B \in \pi} I_B \right)_{t_1 \dots t_r} = \sum_{\rho \leq \pi^*} c(\rho) = (-1)^{\nu-1} \frac{1}{\binom{n-1}{\nu-1}} = \Phi_{t_1 \dots t_r},$$

where the first equality follows by the definition of π^* and $\bigotimes_B I_B$, and the second equality follows directly by the Möbius inversion formula on $\mathbf{\Pi}_r$ as given in Lemma S1.

The second claim follows from the fact that

$$\mathbf{k}_r \stackrel{(S12)}{=} \frac{1}{n} \mathbf{Y}' \bullet \Phi = \frac{1}{n} \sum_{\pi \in \mathbf{\Pi}_r} c(\pi) \bigotimes_{B \in \pi} (\mathbf{Y}' \bullet I_B) \stackrel{(S10)}{=} \sum_{\pi \in \mathbf{\Pi}_r} n^{|\pi|-1} c(\pi) \bigotimes_{B \in \pi} \hat{\boldsymbol{\mu}}_B,$$

where $\hat{\boldsymbol{\mu}}_B$ is the symmetric tensor containing all $|B|$ order sample moments among the variables in B . \square

In the analysis of the asymptotic difference between \mathbf{k}_r and the plug-in estimator $\hat{\mathbf{k}}_r$ we will use the following lemma.

Lemma S5. *For every $\pi \in \Pi_r$ we have*

$$n^{|\pi|-1}c(\pi) - \mathbf{m}(\pi, \mathbf{1}) = O(n^{-1}).$$

Proof. As we noted in the proof of Lemma S4, the Möbius inversion formula in Lemma S1 gives that

$$(S14) \quad \sum_{\rho \leq \pi} c(\rho) = (-1)^{|\pi|-1} \frac{1}{\binom{n-1}{n-|\pi|}}.$$

Let $\mathbf{0} \in \Pi_r$ be the minimal partition into r singleton blocks. By (S14), applied to $\pi = \mathbf{0}$,

$$n^{r-1}c(\mathbf{0}) = (-1)^{r-1} \frac{n^{r-1}}{\binom{n-1}{n-r}} = \mathbf{m}(\mathbf{0}, \mathbf{1}) \frac{n^r}{(n)_r},$$

where $(n)_r = n \cdots (n-r+1)$ is the corresponding falling factorial. In particular, $n^{r-1}c(\mathbf{0}) = \mathbf{m}(\mathbf{0}, \mathbf{1}) + O(n^{-1})$. Now suppose the claim is proven for all partitions with more than l blocks. Let π be a partition with exactly l blocks. If $\rho < \pi$ then $|\rho| > l$ and $n^{|\rho|-1}c(\rho) = \mathbf{m}(\rho, \mathbf{1}) + O(n^{-1})$ so

$$n^{l-1}c(\rho) = n^{l-|\rho|}n^{|\rho|-1}c(\rho) = n^{l-|\rho|}\mathbf{m}(\rho, \mathbf{1}) + O(n^{l-|\rho|-1}) = O(n^{l-|\rho|}).$$

This assures that

$$n^{l-1} \sum_{\rho \leq \pi} c(\rho) = n^{l-1}c(\pi) + O(n^{-1}).$$

Using (S14) in the same way as above, we get that $n^{|\pi|-1}c(\pi) = \mathbf{m}(\pi, \mathbf{1}) + O(n^{-1})$ and now the result follows by recursion. \square

S3.5. Vectorizations of tensors. The dimension of the space of symmetric tensors $S^r(\mathbb{R}^d)$ is $\binom{d+r-1}{r}$. Like for symmetric matrices, it is often convenient to view $T \in S^r(\mathbb{R}^d)$ as a general tensor in $\mathbb{R}^{d \times \cdots \times d}$. In this case $\text{vec}(T) \in \mathbb{R}^{d^r}$ is a vector obtained from all the entries of T .

Throughout the paper we largely avoided vectorization. This operation is however hard to circumvent in the asymptotic considerations. If we make a specific claim about the joint Gaussianity of the entries of a random tensor T , we could use a more invariant approach of Eaton [2007]. However, using vectorizations, makes the calculations more direct without referring to abstract linear algebra.

In this context we also often rely on the matrix-vector version of the tensor equation $S = A \bullet T$

$$(S15) \quad \text{vec}(S) = A^{\otimes r} \cdot \text{vec}(T),$$

where $A^{\otimes r} = A \otimes \cdots \otimes A$ if the r -th Kronecker power of A .

S3.6. Asymptotic distribution of sample statistics. To derive the asymptotic distribution of the minimum distance estimators in Section 6 we require the asymptotic distribution of the sample moments or the \mathbf{k} -statistics.

Specifically, we need the joint distribution of the sample moments/cumulants that are restricted to zero. To derive these in a convenient way we define $m_{S,T} : \mathbb{R}^{d \times d} \rightarrow S^2(\mathbb{R}^d) \oplus S^r(\mathbb{R}^d)$ to be

$$(S16) \quad m_{S,T}(A) = (A \bullet S - I_d, A \bullet T) .$$

The cases that we consider are $S = h_2(Y)$, $T = h_r(Y)$, in which case we write simply $m(A)$, and $S = \hat{h}_2$, $T = \hat{h}_r$, in which case we write $\hat{m}_n(A)$. Here, \hat{h}_r denotes either the sample moments, denoted by $\hat{\mu}_r$, or the r th order \mathbf{k} -statistic, denoted by \mathbf{k}_r , which are computed from a given sample $\{Y_s\}_{s=1}^n$ as discussed above. It is worth pointing out that these results generalize existing results [e.g. Jammalamadaka et al., 2021] for the asymptotic analysis of cumulant estimates to higher order tensors.

Sample moments

The sample moments of Y are defined as in (S9). When using moments the distance measure $\hat{m}_n(A)$ (see (S16)) depends on the tensors $\hat{\mu}_2$ and $\hat{\mu}_r$. As formalized in the lemma below, we have that under suitable moment assumptions that

$$(S17) \quad \hat{\mu}_p \xrightarrow{p} \mu_p(Y) \quad \forall p \leq r ,$$

and

$$(S18) \quad \sqrt{n} \text{vec}(\hat{\mu}_2 - \mu_2(Y), \hat{\mu}_r - \mu_r(Y)) \xrightarrow{d} N(0, V) ,$$

where V is the asymptotic variance matrix with entries

$$V_{i,j} = \text{cov}(Y_{i_1} \cdots Y_{i_k}, Y_{j_1} \cdots Y_{j_l}) \quad k, l \in \{2, r\} .$$

We note that V is not positive definite as vectorizing the tensors does not imply that the entries are unique. We will correct for this when required below. Further V can be consistently estimated by its sample version.

Given (S18) we can use (S15) to derive the limiting distribution of $\hat{m}_n(A)$ for moments. We have

$$(S19) \quad \begin{aligned} \sqrt{n} \text{vec}(\hat{m}_n(A) - m(A)) &= [A^{\otimes 2}, A^{\otimes r}] \cdot \sqrt{n} \text{vec}(\hat{\mu}_2 - \mu_2(Y), \hat{\mu}_r - \mu_r(Y)) \\ &\xrightarrow{d} N(0, A^{2,r} V A^{2,r'}) \end{aligned}$$

where $A^{2,r} = [A^{\otimes 2}, A^{\otimes r}]$. Let the asymptotic variance matrix be denoted by

$$(S20) \quad \Sigma_\mu^{2,r} = A^{2,r} V A^{2,r'} .$$

\mathbf{k} -statistics

Next, we provide analog steps for the \mathbf{k} -statistics. First, let $\mu_{\leq r}$ be the vector containing all moments of a random vector Y of order up to r (it has dimension $\binom{d+r}{r}$). Formula (S7) gives an explicit function for $\kappa_r(Y)$

in terms of $\boldsymbol{\mu}_{\leq r}$. For the vectorized tensor $\kappa_r(Y)$ we define the Jacobian $F = \nabla_{\boldsymbol{\mu}'_{\leq r}} \text{vec}(\kappa_r(Y))$, which is a $d^r \times \binom{d+r}{r}$ matrix. This matrix is not a full rank but only because $\kappa_r(Y)$ is a symmetric tensor which has many repeated entries. The submatrix obtained from F by taking the rows corresponding to the unique entries of $\kappa_r(Y)$ has full row rank. This follows because for any two r -tuples $1 \leq i_1 \leq \dots \leq i_r \leq d$ and $1 \leq j_1 \leq \dots \leq j_r \leq d$ we have that

$$\frac{\partial \kappa_{i_1 \dots i_r}}{\partial \mu_{j_1 \dots j_r}} = \begin{cases} 1 & \text{if } (i_1, \dots, i_r) = (j_1, \dots, j_r), \\ 0 & \text{otherwise,} \end{cases}$$

and so, this submatrix contains the identity matrix.

Under suitable moment conditions we have

$$\hat{\boldsymbol{\mu}}_{\leq r} \xrightarrow{p} \boldsymbol{\mu}_{\leq r} \quad \text{and} \quad \sqrt{n}(\hat{\boldsymbol{\mu}}_{\leq r} - \boldsymbol{\mu}_{\leq r}) \xrightarrow{d} N(0, H)$$

and since $\boldsymbol{\mu}_{\leq r}$ only includes unique moments we may conclude that H is positive definite.

As in Appendix S3.4, denote $\hat{\boldsymbol{\kappa}}_r$ to be the image of $\hat{\boldsymbol{\mu}}_{\leq r}$ under the map (S7). It then follows from the delta method that

$$(S21) \quad \sqrt{n} \text{vec}(\hat{\boldsymbol{\kappa}}_r - \kappa_r(Y)) \xrightarrow{d} N(0, FHF') .$$

We emphasize that this particular estimator of cumulants will not be of direct interest. What we need is the form of the covariance matrix in (S21). We will show that k -statistics \mathbf{k}_r have the same asymptotic distribution.

Lemma S6. *If $\mathbb{E}\|Y_s\|^{2r} < \infty$ we have that*

$$\sqrt{n} \text{vec}(\mathbf{k}_r - \kappa_r(Y)) \xrightarrow{d} N(0, FHF') .$$

Proof. By (S21) and Slutsky lemma, it is enough to show that $\sqrt{n}(\mathbf{k}_r - \hat{\boldsymbol{\kappa}}_r) \xrightarrow{p} 0$. By Lemma S4,

$$[\mathbf{k}_r - \hat{\boldsymbol{\kappa}}_r]_{i_1 \dots i_r} = \sum_{\pi \in \Pi_r} (n^{|\pi|-1} c(\pi) - \mathbf{m}(\pi, \mathbf{1})) \prod_{B \in \pi} \hat{\boldsymbol{\mu}}_{i_B},$$

where the coefficients $c(\pi)$ are defined in (S13). By Lemma S5, $n^{|\pi|-1} c(\pi) - \mathbf{m}(\pi, \mathbf{1}) = O(n^{-1})$ for all $\pi \in \Pi_r$ and so in particular

$$\sqrt{n}(n^{|\pi|-1} c(\pi) - \mathbf{m}(\pi, \mathbf{1})) = o(1).$$

Under the stated moment assumption $\hat{\boldsymbol{\mu}}_{i_B} = O_p(1)$ and so $[\mathbf{k}_r - \hat{\boldsymbol{\kappa}}_r]_{i_1 \dots i_r} = o_p(1)$, which completes the proof. \square

By Lemma S6, every linear transformation of $\sqrt{n} \text{vec}(\mathbf{k}_r - \kappa_r(Y))$ will be also Gaussian. We will be in particular interested in transformations $A^{\otimes r} \text{vec}(\mathbf{k}_r - \kappa_r(Y))$ as motivated by the multilinear action of A on $S^r(\mathbb{R}^d)$ (cf. (S15)). We have

$$\sqrt{n} A^{\otimes r} \text{vec}(\mathbf{k}_r - \kappa_r(Y)) \xrightarrow{d} N(0, A^{\otimes r} F H (A^{\otimes r} F)') .$$

A similar analysis can be given if $\kappa_r(Y)$ is complemented with some other lower order cumulants. We will use one version of that. Let $F^{2,r}$ be the Jacobian matrix of the transformation from $\boldsymbol{\mu}_{\leq r}$ to cumulants $\text{vec}(\kappa_2(Y), \kappa_r(Y)) \in \mathbb{R}^{d^2+d^r}$. By exactly the same arguments as above we get

$$(S22) \quad \sqrt{n} \text{vec}(\mathbf{k}_2 - \kappa_2(Y), \mathbf{k}_r - \kappa_r(Y)) \xrightarrow{d} N(0, F^{2,r} H(F^{2,r})') .$$

Recall from (S16) that $m_{S,T}(A) = (A \bullet S - I_d, A \bullet T)$ and consider $m(A)$ and $\hat{m}_n(A)$ as defined by cumulants and \mathbf{k} -statistics in Section 6.

$$(S23) \quad \text{vec}(\hat{m}_n(A) - m(A)) = [A^{\otimes 2}, A^{\otimes r}] \cdot \text{vec}(\mathbf{k}_2 - \kappa_2(Y), \mathbf{k}_r - \kappa_r(Y)) .$$

We will write $A^{2,r} = [A^{\otimes 2}, A^{\otimes r}]$ and, using (S22), we immediately conclude

$$\sqrt{n} \text{vec}(\hat{m}_n(A) - m(A)) \xrightarrow{d} N(0, A^{2,r} F^{2,r} H(A^{2,r} F^{2,r})') .$$

Let this asymptotic covariance matrix be denoted by

$$(S24) \quad \Sigma_{\mathbf{k}}^{2,r} = A^{2,r} F^{2,r} H(A^{2,r} F^{2,r})' .$$

We summarize these general results in the following lemma adopting the notation required for the main text.

Lemma S7. *Suppose $\{Y_s\}_{s=1}^n$ is i.i.d.*

- (1) *if $\mathbb{E}\|Y_s\|^r < \infty$, then $\hat{\boldsymbol{\mu}}_p - \mu_p(Y) \xrightarrow{p} 0$ and $\mathbf{k}_p - \kappa_p(Y) \xrightarrow{p} 0$ for all $p \leq r$.*
- (2) *if $\mathbb{E}\|Y_s\|^{2r} < \infty$, then*

$$\sqrt{n} \text{vec}(\hat{m}_n(A) - m(A)) \xrightarrow{d} N(0, \Sigma_h^{2,r}) \quad h = \mu, \mathbf{k} ,$$

where $h = \mu$ or $h = \mathbf{k}$ depends on whether $\hat{m}_n(A)$ and $m(A)$ are based on moments or cumulants, respectively. We have that the moment based variance $\Sigma_{\mu}^{2,r}$ is defined in (S20) and the cumulant based variance $\Sigma_{\mathbf{k}}^{2,r}$ in (S24).

S4. ADDITIONAL INFERENCE TOOLS

In this section we complement the inference Section 6 with some additional tools that can be used to select the appropriate moment/cumulant zero restrictions in a data driven way.

S4.1. Testing over-identifying restrictions. While zero restrictions on higher order moments or cumulants can be motivated from several angles (cf. the discussion in Section 2), it is useful to test ex-post whether the restrictions indeed appear to hold in a given application. In the setting where d_g is strictly greater than d^2 , i.e. the total number of restrictions is larger when compared to the number of parameters in A , we can conduct a general specification test following the approach outlined in Hansen [1982].

Proposition S1. *If the conditions of Proposition 6.3 hold we have that as $n \rightarrow \infty$*

$$\Lambda_n := n \hat{L}_{\hat{\Sigma}_n^{-1}}(\hat{A}_{\hat{\Sigma}_n^{-1}}) \xrightarrow{d} \chi^2(d_g - d^2) .$$

The proposition implies that Λ_n can be viewed as a test statistic for verifying the identifying restrictions. Specifically, when $g(QA_0) \neq 0$ the statistic Λ_n diverges under most alternatives. That said, if any of the other assumptions fails, e.g. the moment condition, the statistic will also fail to converge to a $\chi^2(d_g - d^2)$ random variable. This implies that we should view Proposition S1 as a general test for model misspecification.

A more refined test can be formulated when sufficient confidence exists in a subset of the identifying restrictions. To set this up let $g(A) = (g_1(A), g_2(A))$ be a partition of the identifying moment/cumulant restrictions such that $g_1(A)$ has dimension $d_{g_1} \geq d^2$. We propose a test for whether the additional identifying restrictions $g_2(A)$ are valid.

Denote as earlier $\Lambda_n = n\hat{L}_{\hat{\Sigma}_n^{-1}}(\hat{A}_{\hat{\Sigma}_n^{-1}})$ and let Λ_n^0 be similarly defined by for a smaller set of identifying restrictions.

Proposition S2. *If the conditions of Proposition 6.3 hold we have that as $n \rightarrow \infty$*

$$C_n := \Lambda_n - \Lambda_n^0 \xrightarrow{d} \chi^2(d_g - d_{g_1}) .$$

The test statistic C_n allows to verify whether adding the additional identifying restrictions $g_2(A)$ is valid. The test rejects when $g_2(QA_0) \neq 0$, that is, when the additional restrictions do not hold.

S5. COMPUTING THE ASYMPTOTIC VARIANCE

In this section we give computational details for estimating the asymptotic variance matrices Σ and S as defined in Proposition 6.3. Starting with Σ (see equation (19)) we first recall that Σ is really Σ_h and the expression depends on whether moment or cumulant restrictions are used. For moments we obtained

$$\Sigma_\mu = D_{\mathcal{I}}^{2,r} \Sigma_\mu^{2,r} D_{\mathcal{I}}^{2,r'} \quad \text{with} \quad \Sigma_\mu^{2,r} = A^{2,r} V A^{2,r'} ,$$

and for cumulants

$$\Sigma_\kappa = D_{\mathcal{I}}^{2,r} \Sigma_\kappa^{2,r} D_{\mathcal{I}}^{2,r'} \quad \text{with} \quad \Sigma_\kappa^{2,r} = A^{2,r} F^{2,r} H (A^{2,r} F^{2,r})' ,$$

where $D_{\mathcal{I}}^{2,r}$ is a selection matrix that selects the corresponding to the unique entries in $S^r(\mathbb{R}^d) \oplus \mathcal{V}^\perp$, V and H contain the covariances of $\text{vec}(\hat{\mu}_2, \hat{\mu}_r)$ and $\hat{\mu}_{\leq r}$, respectively, $A^{2,r} = [A^{\otimes 2}, A^{\otimes r}]$ and $F^{2,r}$ is the Jacobian matrix of the transformation from $\mu_{\leq r}$ to cumulants (κ_2, κ_r) , see Section S3.6 for explicit definitions.

The moment matrices V and H and the Jacobian matrix $F^{2,r}$ can be estimated by replacing the population moments of $\mu_r(Y)$ by the sample moments $\hat{\mu}_r$. Further, $A^{2,r} = [A^{\otimes 2}, A^{\otimes r}]$ can be replaced by its estimate $\hat{A}_{W_n}^{2,r} = [\hat{A}_{W_n}^{\otimes 2}, \hat{A}_{W_n}^{\otimes r}]$. Combining we obtain the estimates

$$\hat{\Sigma}_{\mu,n} = D_{\mathcal{I}}^{2,r} \hat{V} D_{\mathcal{I}}^{2,r'} \quad \text{and} \quad \hat{\Sigma}_{\kappa,n} = D_{\mathcal{I}}^{2,r} \hat{A}_{W_n}^{2,r} \hat{F}^{2,r} \hat{H} (\hat{A}_{W_n}^{2,r} \hat{F}^{2,r})' D_{\mathcal{I}}^{2,r'} .$$

While these plug-in estimators are conceptually straightforward, for cumulants it does require determining the Jacobian $F^{2,r}$, which can be a tedious task.

Therefore, for cumulant restriction we recommend estimating Σ_h using a simple bootstrap. Let $\hat{\varepsilon}_n = \hat{A}_{W_n} \mathbf{Y}_n$ denote the $n \times 1$ vector of residuals. We can resample these residuals (with replacement) to get $\hat{\varepsilon}_n^*$ and construct bootstrap draws of $\hat{g}_n(\hat{A}_{W_n})$, say g_n^* . Repeating this B times allows to compute the bootstrap variance estimate

$$\hat{\Sigma}_n/n = \frac{1}{B} \sum_{b=1}^B g_n^{*,b} g_n^{*,b'}.$$

The $1/n$ comes from the definition $\Sigma = \lim_{n \rightarrow \infty} \text{var}(\sqrt{n} \hat{g}_n(QA_0))$. Using the bootstrap has the benefit that no additional analytical calculations are needed and evaluating $g_n^{*,b}$ only requires computing the sample statistics $\mu_p(Y)$ or k_p , for $p = 2, r$, for each bootstrap draw $\hat{\varepsilon}_n^*$.

While the bootstrap is conceptually attractive, it is worth noting that, at least in principle, the covariance between two \mathbf{k} -statistics $\mathbf{k}_{i_1 \dots i_r}$ and $\mathbf{k}_{j_1 \dots j_r}$ can be computed exactly for any given sample size using the general formula for cumulants of \mathbf{k} -statistics as given in Section 4.2.3 in [McCullagh \[2018\]](#). Although the covariance is arguably the simplest cumulant, the formula still involves combinatorial quantities that are hard to obtain. Given the moments of Y , we could also use the explicit formula (S12) to obtain the covariance in any given case by noting that

$$\mathbb{E} \text{vec}(\mathbf{k}_r) \text{vec}(\mathbf{k}_r)' = \frac{1}{n^2} \mathbb{E} [(\mathbf{Y}')^{\otimes r} \text{vec}(\Phi) \text{vec}(\Phi)' \mathbf{Y}^{\otimes r}].$$

Note however that $\text{vec}(\Phi)$ has n^r entries with many of them repeated, so the naive approach is very inefficient. An efficient, perhaps umbral, approach to these symbolic computations could help to obtain better estimates of A .

Next, we compute the asymptotic variance $S = (G' \Sigma^{-1} G)^{-1}$, where $G = G(QA_0)$ is the Jacobian matrix corresponding to $g(A)$. Combining the estimator \hat{A}_{W_n} and the map (S28) provides the estimate for G . Combining this an estimate for Σ as defined above allows to estimate S .

S6. ADDITIONAL NUMERICAL RESULTS

In this section we provide additional simulation results that complement Section 7. We compare the performance of the minimum distance estimators across different measures, dimensions and sample sizes.

S6.1. Alternative performance measures. We start by providing the same results as in the main text but now measuring the accuracy of the different procedures in terms of the Frobenius distance d_F [e.g. [Chen and Bickel, 2006](#)] which is often referred to as the minimum distance index [e.g.

[Ilmonen et al., 2010](#)] and can be defined as

$$d_F(\hat{A}_{W_n}, A_0) = \min_{Q \in SP(d)} \frac{1}{d^2} \|\hat{A}_{W_n}^{-1} Q A_0 - I_d\|_F,$$

where the scaling by d^2 is an arbitrary choice.

For this distance measure Tables [S1](#) and [S2](#) replicate Tables [2](#) and [3](#) from the main text. We find that the minimum distance estimator based on the reflectionally invariant restrictions remains to perform well across all specifications. Also for skewed densities the minimum distance estimator based on the diagonal third order tensor restrictions performs well.

For the common variance model, i.e. Table [S1](#), there are a few differences with respect to the Amari errors that are worth pointing out. First, TICA performs relatively less well. Further inspection showed that is largely due to the small sample size and the performance of TICA improves considerably when n increases. Second, some of the ICA methods (e.g. FastICA and JADE) perform well for $t(5)$, SKU and KU densities.

For the multiple scaled elliptical model, i.e. Table [S2](#), the results are very similar when compared to the Amari errors, and the minimum distance estimator based on the reflectionally invariant restrictions is always preferred.

S6.2. Larger experiments common variance model. In the main text in Figure [1](#) we showed the results for the common variance model with $d = 5$ and $n = 200, 1000$ corresponding to two specific distributions for the errors η_i : the $t(5)$ distribution as well as the Bi-Modal distribution BM . Here we show the same results but also include the other densities from Table [1](#).

Specifically, Figures [S1-S3](#) show all experiments that we conducted for the common variance model. The following additional results are worth mentioning. First, when the true errors correspond to the normal distribution the variances of all estimators are large and do not shrink noticeably when n increases. The reason under normal errors for η_i the deviations from the Gaussian distribution of $\varepsilon_i = \tau \eta_i$, with independent $\tau \sim \text{gamma}(1, 1)$, is very close to the Gaussian distribution and hence the parameters are poorly identified.

Second, for $n = 1000$ we find that the diagonal tensor restrictions based on the third moments work well for the Skewed Unimodal (SKU) density. For $n = 200$ the evidence is not convincing, but for larger sample sizes these restrictions in combination with the efficient weighting matrix yield good performance. Only TICA, which assumes that K is known, leads to better performance.

Third, in general TICA works well for Student's t type densities like, $t(5)$ and Kurtotic Unimodal (KU). The reason is that, in addition to exploiting knowledge of K , the objective function of TICA is close in shape to the Student's t density [see [Hyvärinen et al., 2001](#), equation 3.10]. As such TICA behaves like the MLE estimator for these densities.

Fourth, for all other densities which impose larger deviations from the Student t shape the estimators that were based on the reflectional invariant restrictions always perform better. The benefits are most clearly shown for bi modal densities.

Fifth, using the efficient weighting matrix shows most advantages for large sample sizes. The reason is that estimating the efficient weighting matrix accurately requires a large sample size. This improvement in weighting matrix accuracy is directly reflected in the Amari errors.

S6.3. Larger experiments multiple scaled elliptical. Next, we revisit the nICA model with multiple scaled elliptical errors as presented in (4). For this model comparative simulation results were shown in Section 7.2 for $d = 2$ and $n = 200$. Here we consider the specifications where $d = 5$ and $n = 200, 1000$. Figures S4-S6 show the results. Overall, the results for the scaled elliptical components model are quite similar across the densities for η_i . As in Table 3 the means of the Amari errors are roughly equal, but there exist some variations in the variances. First, except for the Gaussian density (which does not yield an identified model) when n increases the variances generally decrease. Second, the evidence in favor of the efficient weighting matrix is mixed often the identity weighting matrix is preferred. This is most likely due to the fact that the multiple scaled elliptical model has quite heavy tails which may invalidate the moment assumptions needed for the consistent estimation of the weighting matrix, or at least reduce the accuracy of the weighting matrix estimate.

TABLE S1. MINIMUM DISTANCE: COMMON VARIANCE MODEL

Non-Independent Components Analysis										
Method	\mathcal{N}	$t(5)$	SKU	KU	BM	SBM	SKB	TRI	CL	ACL
$\mu_3^{d,I}$	0.16	0.13	0.13	0.13	0.18	0.20	0.16	0.18	0.16	0.15
$\mu_3^{d,\widehat{\Sigma}_n^{-1}}$	0.14	0.12	0.11	0.12	0.15	0.17	0.13	0.15	0.13	0.13
$\mu_4^{r,I}$	0.11	0.12	0.12	0.11	0.10	0.08	0.11	0.10	0.12	0.11
$\mu_4^{r,\widehat{\Sigma}_n^{-1}}$	0.11	0.11	0.11	0.10	0.09	0.05	0.10	0.08	0.10	0.11
TICA	0.19	0.18	0.18	0.18	0.23	0.25	0.22	0.24	0.20	0.20
Independent Components Analysis										
Method	\mathcal{N}	$t(5)$	SKU	KU	BM	SBM	SKB	TRI	CL	ACL
Fast	0.14	0.09	0.11	0.08	0.20	0.25	0.18	0.21	0.15	0.15
JADE	0.15	0.10	0.11	0.07	0.23	0.26	0.21	0.23	0.18	0.17
Kernel	0.15	0.11	0.12	0.10	0.20	0.25	0.18	0.21	0.16	0.16
ProDen	0.15	0.79	0.30	0.64	0.23	0.60	0.21	0.26	0.17	2.78
Efficient	0.14	0.17	0.16	0.17	0.14	0.15	0.14	0.14	0.14	0.14
NPML	0.14	0.14	0.14	0.14	0.16	0.15	0.15	0.15	0.15	0.15

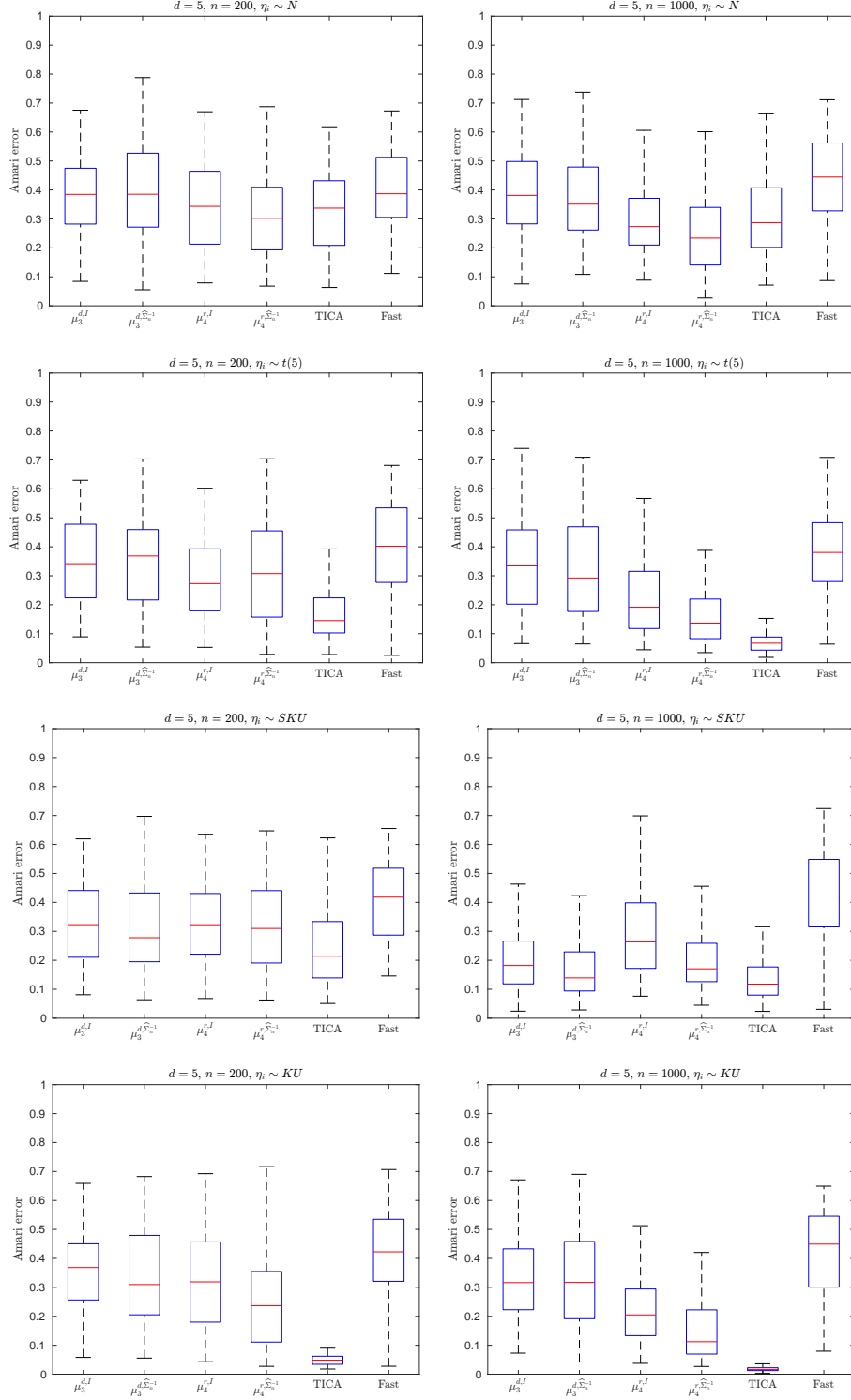
Notes: The table reports the average Minimum Distance Index (across $S = 1000$ simulations) for data sampled from the common variance model (3) with $d = 2$ and $n = 200$. The columns correspond to the different errors considered for the components of η , see Table 1. The top panel reports the errors 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].

TABLE S2. MINIMUM DISTANCE: SCALED ELLIPTICAL

Non-Independent Components Analysis										
Method	\mathcal{N}	$t(5)$	SKU	KU	BM	SBM	SKB	TRI	CL	ACL
$\mu_3^{d,I}$	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
$\mu_3^{d,\widehat{\Sigma}_n^{-1}}$	0.14	0.13	0.13	0.13	0.14	0.14	0.14	0.14	0.13	0.13
$\mu_4^{r,I}$	0.08	0.09	0.09	0.08	0.08	0.08	0.09	0.09	0.09	0.09
$\mu_4^{r,\widehat{\Sigma}_n^{-1}}$	0.09	0.09	0.09	0.09	0.09	0.09	0.08	0.08	0.08	0.08
TICA	0.15	0.16	0.15	0.16	0.14	0.15	0.15	0.15	0.15	0.15
Independent Components Analysis										
Method	\mathcal{N}	$t(5)$	SKU	KU	BM	SBM	SKB	TRI	CL	ACL
Fast	0.14	0.13	0.14	0.14	0.13	0.14	0.14	0.14	0.13	0.14
JADE	0.14	0.14	0.14	0.14	0.14	0.14	0.15	0.14	0.14	0.14
Kernel	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
ProDen	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.15
Efficient	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
NPML	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14

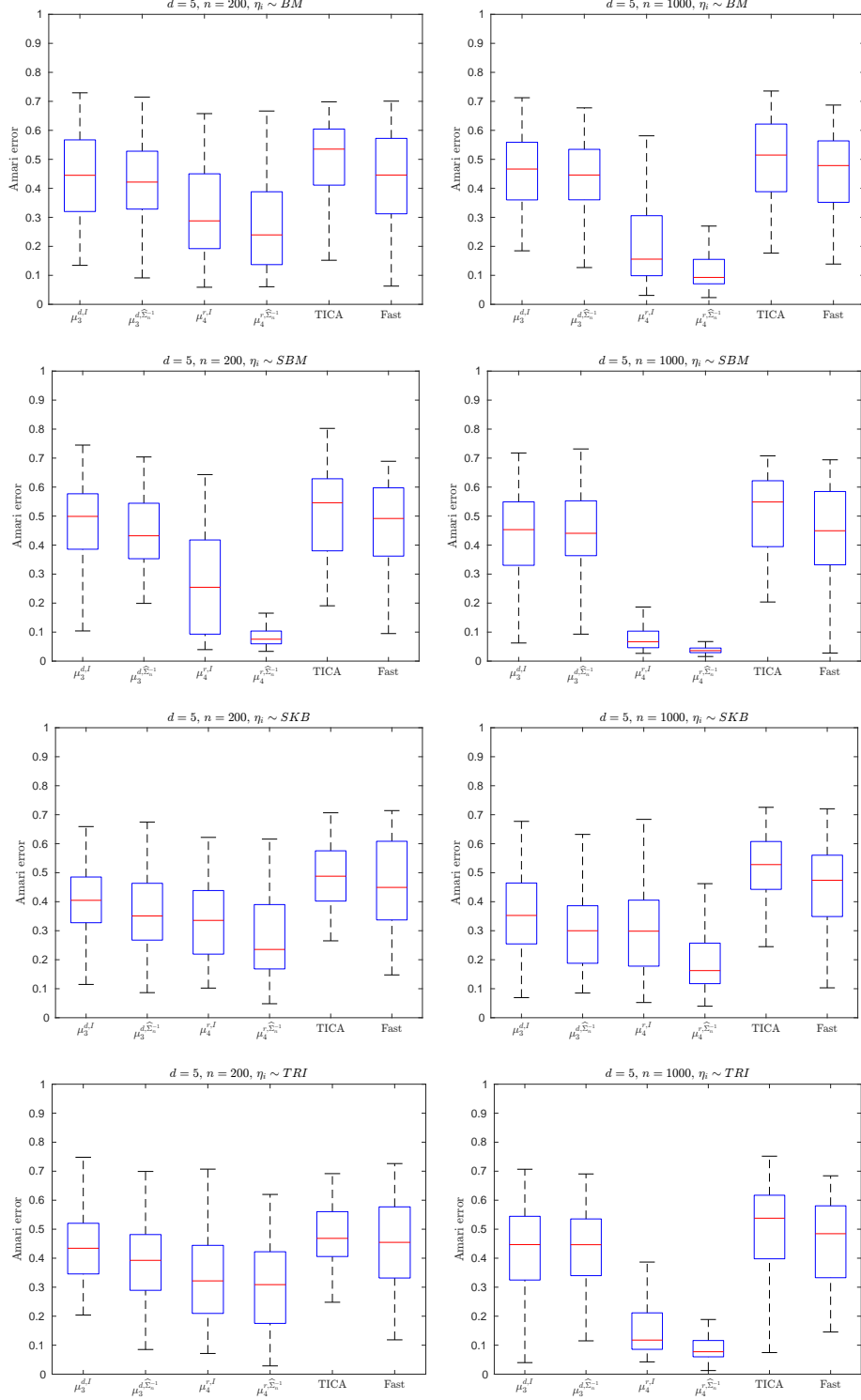
Notes: The table reports the average Minimum Distance errors (across $S = 1000$ simulations) for data sampled from the multiple scaled elliptical model (4) with $d = 2$ and $n = 200$. The columns correspond to the different errors considered for the components of η , see Table 1. The top panel reports 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].

FIGURE S1. COMMON VARIANCE EXPERIMENTS



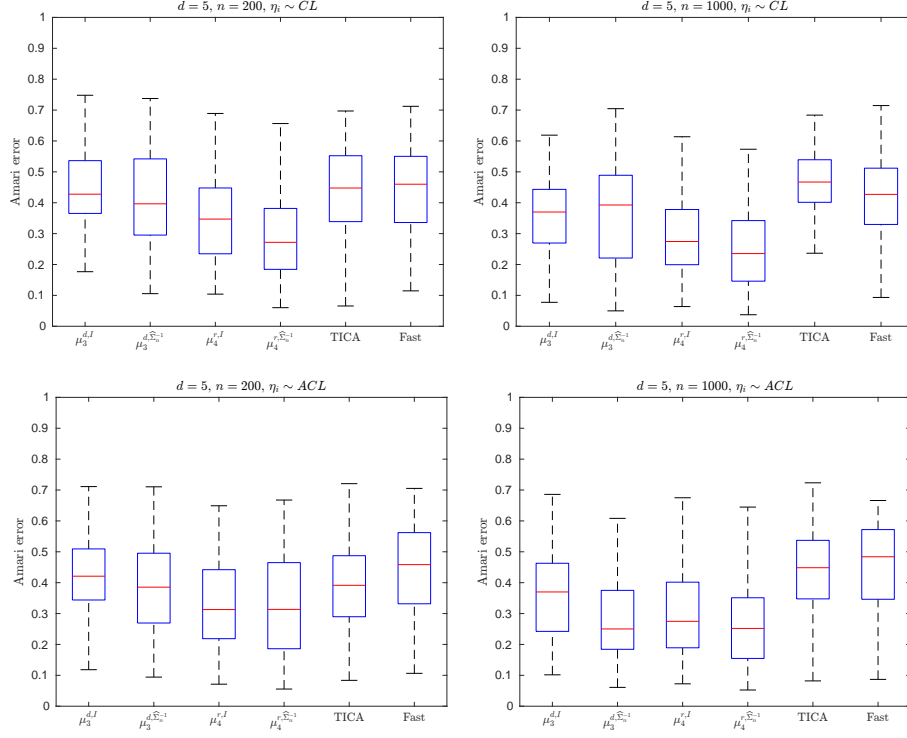
Notes: The figure shows the boxplots for the Amari errors (across $S = 100$ simulations) for data sampled from the common variance model (3). The different settings for the simulations designs are described in the titles and the x -labels indicate the different estimation methods used.

FIGURE S2. COMMON VARIANCE EXPERIMENTS



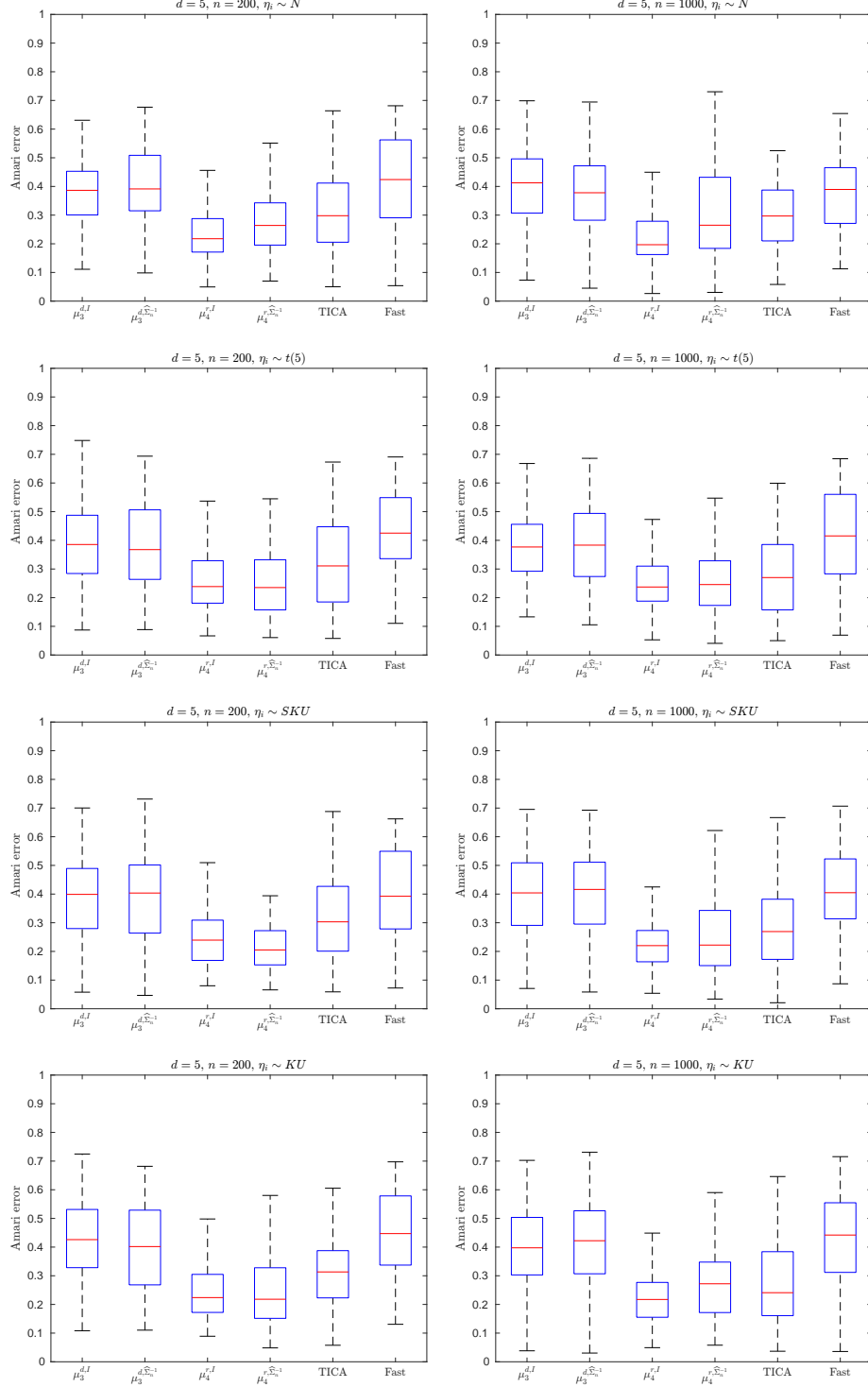
Notes: The figure shows the boxplots for the Amari errors (across $S = 100$ simulations) for data sampled from the common variance model (3). The different settings for the simulations designs are described in the titles and the x -labels indicate the different estimation methods used.

FIGURE S3. COMMON VARIANCE EXPERIMENTS



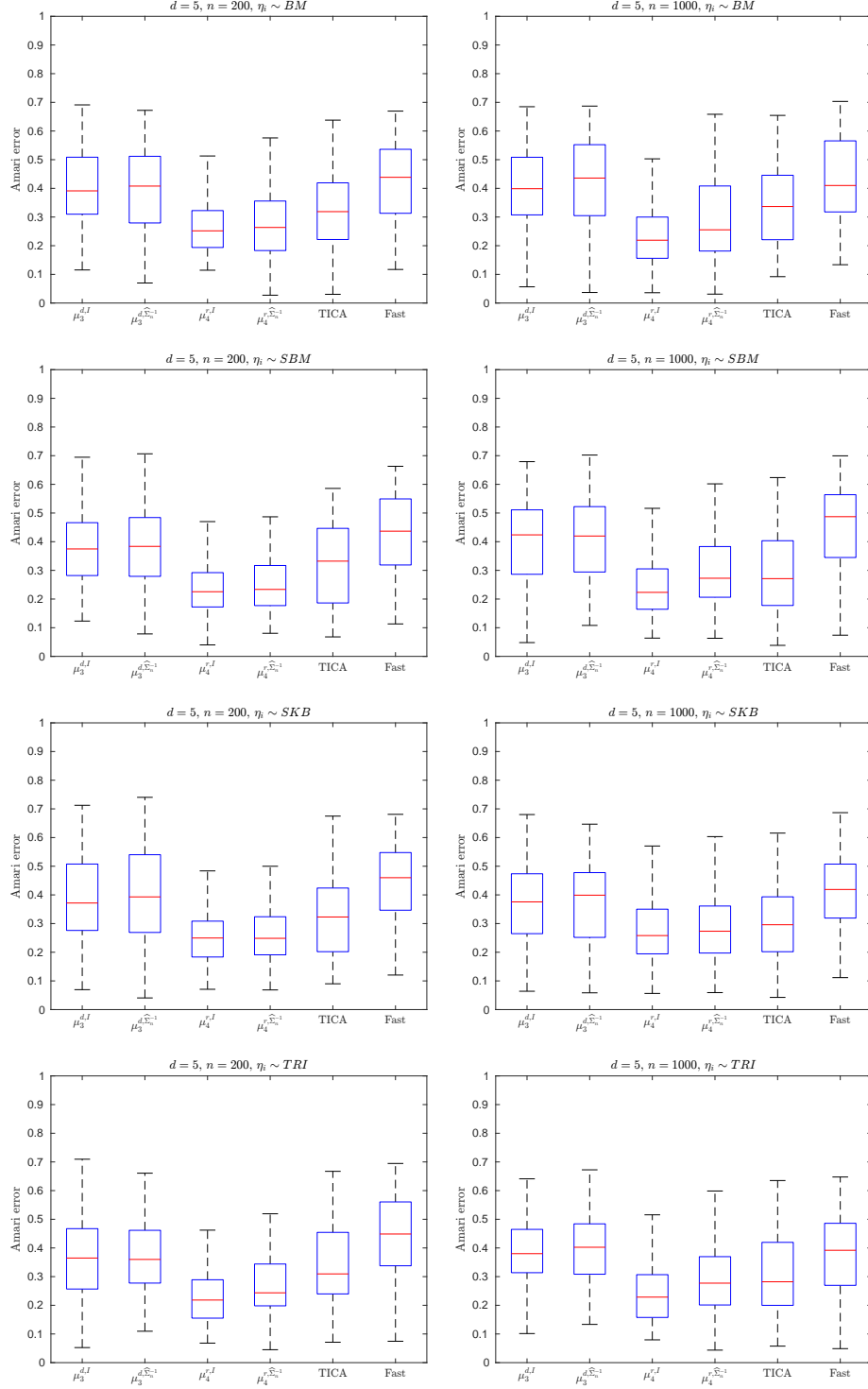
Notes: The figure shows the boxplots for the Amari errors (across $S = 100$ simulations) for data sampled from the common variance model (3). The different settings for the simulations designs are described in the titles and the x -labels indicate the different estimation methods used.

FIGURE S4. SCALED ELLIPTICAL EXPERIMENTS



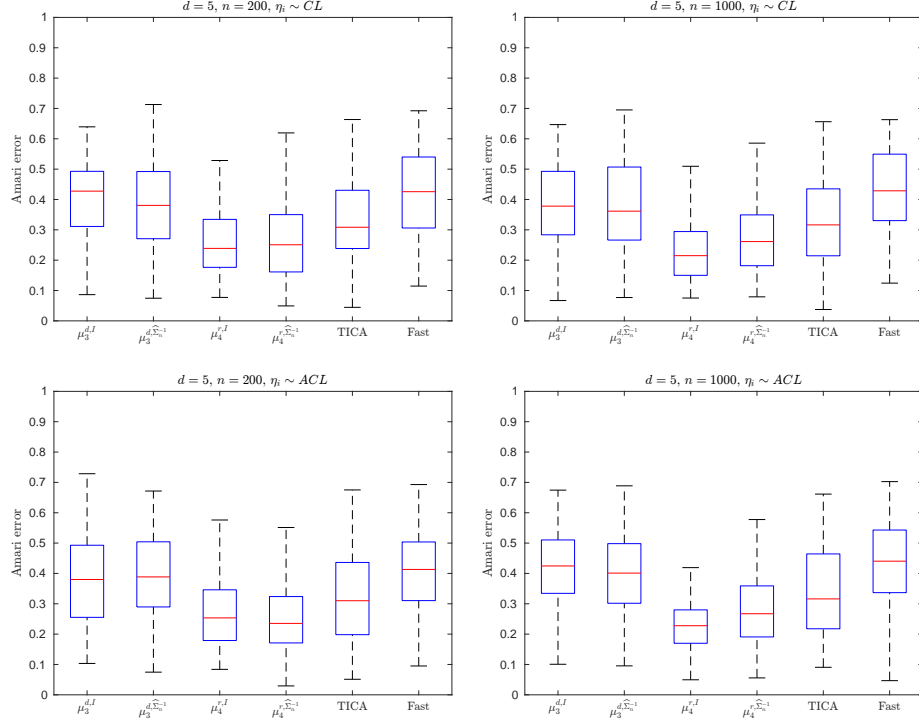
Notes: The figure shows the boxplots for the Amari errors (across $S = 100$ simulations) for data sampled from the multiple scaled elliptical components model (4). The different settings for the simulations designs are described in the titles and the x -labels indicate the different estimation methods used.

FIGURE S5. SCALED ELLIPTICAL EXPERIMENTS



Notes: The figure shows the boxplots for the Amari errors (across $S = 100$ simulations) for data sampled from the multiple scaled elliptical components model (4). The different settings for the simulations designs are described in the titles and the x -labels indicate the different estimation methods used.

FIGURE S6. SCALED ELLIPTICAL EXPERIMENTS



Notes: The figure shows the boxplots for the Amari errors (across $S = 100$ simulations) for data sampled from the multiple scaled elliptical components model (4). The different settings for the simulations designs are described in the titles and the x -labels indicate the different estimation methods used.

S7. OMITTED PROOFS

S7.1. Proof of Proposition 5.17. The condition $Q \bullet T \in \mathcal{V}$ translates into two equations $(Q \bullet T)_{12\dots 2} = (Q \bullet T)_{1\dots 12} = 0$. In other words,

$$Q_{11} \sum_j Q_{2j_1} \cdots Q_{2j_{r-1}} T_{1j} + Q_{12} \sum_j Q_{2j_1} \cdots Q_{2j_{r-1}} T_{2j} = 0$$

and

$$Q_{21} \sum_j Q_{1j_1} \cdots Q_{1j_{r-1}} T_{1j} + Q_{22} \sum_j Q_{1j_1} \cdots Q_{1j_{r-1}} T_{2j} = 0,$$

where in both cases the sum goes over all $(r-1)$ -tuples $\mathbf{j} = (j_1, \dots, j_{r-1}) \in \{1, 2\}^{r-1}$. Note that, since T is symmetric, the entry $T_{\mathbf{i}}$ depends only on how many times 1 appears in \mathbf{i} . Write $t_k = T_{\mathbf{i}}$ if \mathbf{i} has k ones, $k = 0, \dots, r$. With this notation the two equations above simplify to

$$\sum_{k=0}^{r-1} \binom{r-1}{k} Q_{11} Q_{21}^k Q_{22}^{r-1-k} t_{k+1} + \sum_{k=0}^{r-1} \binom{r-1}{k} Q_{12} Q_{21}^k Q_{22}^{r-1-k} t_k = 0$$

and

$$\sum_{k=0}^{r-1} \binom{r-1}{k} Q_{21} Q_{11}^k Q_{12}^{r-1-k} t_{k+1} + \sum_{k=0}^{r-1} \binom{r-1}{k} Q_{22} Q_{11}^k Q_{12}^{r-1-k} t_k = 0.$$

If one of the entries of Q is zero then Q is a permutation matrix. So assume that Q has no zeros. Assume also without loss of generality that Q is a rotation matrix, that is, $Q_{11} = Q_{22}$ and $Q_{12} = -Q_{21}$. Denote $z = Q_{21}/Q_{11}$, which corresponds to the tangent of the rotation angle and so it can take any non-zero value (zero is not possible as $Q_{21} \neq 0$). With this notation and after dividing by Q_{11}^r , the two equations become

$$(S25) \quad \sum_{k=0}^{r-1} \binom{r-1}{k} z^k t_{k+1} - \sum_{k=0}^{r-1} \binom{r-1}{k} z^{k+1} t_k = 0$$

and

$$\sum_{k=0}^{r-1} \binom{r-1}{k} (-1)^{r-1-k} z^{r-k} t_{k+1} + \sum_{k=0}^{r-1} \binom{r-1}{k} (-1)^{r-1-k} z^{r-1-k} t_k = 0.$$

It is convenient to rewrite the latter as

$$(S26) \quad \sum_{k=0}^{r-1} \binom{r-1}{k} (-1)^k z^{k+1} t_{r-k} + \sum_{k=0}^{r-1} \binom{r-1}{k} (-1)^k z^k t_{r-k-1} = 0.$$

Using the fact that $t_1 = t_{r-1} = 0$, (S25) can be written as

$$\sum_{k=1}^{r-1} \left(\binom{r-1}{k} t_{k+1} - \binom{r-1}{k-1} t_{k-1} \right) z^k = 0.$$

and (S26) can be written as

$$\sum_{k=1}^{r-1} \left(\binom{r-1}{k} t_{r-k-1} - \binom{r-1}{k-1} t_{r-k+1} \right) (-z)^k = 0.$$

Since $z \neq 0$, we can divide by it and in both cases we obtain two polynomials of order $r-2$. The first polynomial has coefficients

$$a_k = \binom{r-1}{k+1} t_{k+2} - \binom{r-1}{k} t_k \quad \text{for } k = 0, \dots, r-2$$

and the second has coefficients

$$b_k = (-1)^{k-1} \left(\binom{r-1}{k+1} t_{r-k-2} - \binom{r-1}{k} t_{r-k} \right) = (-1)^k a_{r-k-2}.$$

A common zero for these two polynomials exists if and only if the corresponding resultant is zero. Resultant is defined as the determinant of a certain matrix populated with the coefficients of both polynomials. After reordering the columns of this matrix, we obtain

$$\begin{bmatrix} a_0 & a_{r-2} & 0 & 0 & \cdots & 0 & 0 \\ a_1 & -a_{r-3} & a_0 & a_{r-2} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ a_{r-2} & (-1)^r a_0 & a_{r-3} & (-1)^{r-1} a_1 & \cdots & a_0 & a_{r-2} \\ 0 & 0 & a_{r-2} & (-1)^r a_0 & \cdots & a_1 & -a_{r-2} \\ 0 & 0 & 0 & 0 & \cdots & a_2 & a_{r-3} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & a_{r-2} & (-1)^r a_0 \end{bmatrix}.$$

The first two columns are linearly independent of each other unless the second is a multiple of the first. Indeed, if r is odd, this is only possible if $a_0 = \cdots = a_{r-2} = 0$ (which cannot hold under the genericity assumptions). If r is even this is possible if and only if either $a_k = (-1)^k a_{r-2-k}$ for all k , or $a_k = (-1)^{k-1} a_{r-2-k}$ for all k (which cannot hold under the genericity assumptions). By the same argument, the third and the fourth column are independent of each other and linearly independent of the previous two. Proceeding resursively like that, we conclude that all columns in this matrix are linearly independent proving that the two polynomials cannot have common roots. In other words, there is no rotation matrix apart from the 0° and the 90° rotation matrices that satisfy $Q \bullet T \in \mathcal{V}$.

S7.2. Proof of Proposition 6.2. The proof follows from verifying the conditions for consistency of a general extremum estimator. Specifically, we will verify the conditions of Theorem 2.1 in Newey and McFadden [1994]. We restate the theorem for completeness.

Theorem S1. Suppose that $\hat{\theta}$ minimizes $\hat{L}_n(\theta)$ over $\theta \in \Theta$. Assume that there exists a function $L_0(\theta)$ such that (a) $L_0(\theta)$ is uniquely minimized at θ_0 ,

(b) $L_0(\theta)$ is continuous, (c) Θ is compact and (d) $\sup_{\theta \in \Theta} |\hat{L}_n(\theta) - L_0(\theta)| \xrightarrow{p} 0$, then $\hat{\theta} \xrightarrow{p} \theta_0$.

Next, we verify assumptions (a)-(d) under assumptions (i)-(iv) stated in Proposition 6.2. First, note that \hat{A}_{W_n} minimizes $\hat{L}_{W_n}(A)$ and we take $L_W(A)$ as $L_0(\theta)$ in Theorem S1. Second, in our case the minimizer of $L_W(A)$ is not unique but will correspond to any of the finite points QA_0 for some $Q \in SP(d)$. It follows that our consistency result will only be up to permutation and sign changes of the true A_0 [e.g. Chen and Bickel, 2006]. Formally, for (a): suppose that A is such that $A \neq QA_0$ for any $Q \in SP(d)$, then $g(A) \neq 0$ by assumption (i) and, since W is positive definite by (ii), we have $L_W(A) > 0$. Hence it follows that $L_W(A)$ is only minimized at QA_0 for some $Q \in SP(d)$. Condition (b) follows as $L_W(A)$ is a composition of two polynomial maps. Condition (c) follows from (ii). Condition (d) is assured by the following result.

Lemma S2. Suppose that $\{Y_s\}_{s=1}^n$ is i.i.d, $W_n \xrightarrow{p} W$, $\mathbb{E}\|Y_s\|^r < \infty$, and $\mathcal{A} \subset GL(d)$ is a compact set. Then

$$\sup_{A \in \mathcal{A}} |\hat{L}_{W_n}(A) - L_W(A)| \xrightarrow{p} 0$$

Proof. First, note that given the i.i.d. assumption and the moment condition (iv) we have that $\|\hat{\mu}_p - \mu_p(Y)\| \xrightarrow{p} 0$ and $\|\mathbf{k}_p - \kappa_r(Y)\| \xrightarrow{p} 0$ for any $p \leq r$ by Lemma S7 part 1. Note that the norm $\|\cdot\|$ on the tensor is defined in the usual way as the sum of the squares of all elements. Using the general notation of Section 6 we have that $\|\hat{h}_p - h_p(Y)\| \xrightarrow{p} 0$ for $p \leq r$. Hence,

$$\sup_{A \in \mathcal{A}} \|A^{\otimes p} \text{vec}(\hat{h}_p - h_p(Y))\|^2 \leq \|\hat{h}_p - h_p(Y)\|^2 \sup_{A \in \mathcal{A}} \|A^{\otimes p}\|^2 \xrightarrow{p} 0.$$

Here we used the fact that \mathcal{A} is a compact and so, in particular, $\|A^{\otimes p}\|^2$ is bounded on \mathcal{A} .

Using (S23), we get

$$\begin{aligned} \sup_{A \in \mathcal{A}} \|\hat{m}_n(A) - m(A)\|^2 &\leq \sup_{A \in \mathcal{A}} \|A^{\otimes 2} \text{vec}(\hat{h}_2 - h_2(Y))\|^2 \\ &\quad + \sup_{A \in \mathcal{A}} \|A^{\otimes r} \text{vec}(\hat{h}_r - h_r(Y))\|^2 \xrightarrow{p} 0. \end{aligned}$$

As $g_{S,T}(A)$ is defined in (16) as a projection of $m_{S,T}(A)$ on certain coordinates, we conclude that

$$\sup_{A \in \mathcal{A}} \|\hat{g}_n(A) - g(A)\| \xrightarrow{p} 0.$$

By the triangle inequality

$$\left| \hat{L}_{W_n}(A) - L_W(A) \right| \leq \left| \hat{L}_{W_n}(A) - L_{W_n}(A) \right| + |L_{W_n}(A) - L_W(A)|.$$

The second term is readily bounded by $\|g(A)\|^2 \|W_n - W\|$ using the basic operator norm inequality. To bound the first term, note that, by the triangle

inequality

$$\left| \hat{L}_{W_n}(A) - L_{W_n}(A) \right| = \left| \|\hat{g}_n(A)\|_{W_n}^2 - \|g(A)\|_{W_n}^2 \right| \leq \|\hat{g}_n(A) - g(A)\|_{W_n}^2,$$

which can be bounded by $\|\hat{g}_n(A) - g(A)\|^2 \|W_n\|$. We conclude that

$$\left| \hat{L}_{W_n}(A) - L_W(A) \right| \leq \|\hat{g}_n(A) - g(A)\|^2 \|W_n\| + \|g(A)\|^2 \|W_n - W\|.$$

It follows that $\sup_{A \in \mathcal{A}} |\hat{L}_{W_n}(A) - L_W(A)| \xrightarrow{P} 0$ as required. \square

We may now apply Theorem S1 to conclude that $\hat{A}_{W_n} \xrightarrow{P} QA_0$ for some $Q \in \text{SP}(d)$.

S7.3. Proof of Proposition 6.3. The proof follows from verifying the conditions for asymptotic normality of a generalized moment or distance estimator. Specifically, we will verify the conditions of Theorem 3.2 in Newey and McFadden [1994]. We restate the theorem for completeness.

Theorem S3. *Suppose that $\hat{\theta}$ minimizes $\hat{L}_n(\theta)$ over $\theta \in \Theta$ with Θ compact, where $\hat{L}_n(\theta)$ is of the form $\hat{g}_n(\theta)'W_n\hat{g}_n(\theta)$ and $W_n \xrightarrow{P} W$ with W positive semi-definite, $\hat{\theta} \xrightarrow{P} \theta_0$ and (a) $\theta_0 \in \text{Int}(\Theta)$, (b) $\hat{g}_n(\theta)$ is continuously differentiable in a neighborhood \mathcal{N} of θ_0 , (c) $\sqrt{n}\hat{g}_n(\theta_0) \xrightarrow{d} N(0, \Omega)$, (d) there is $G(\theta)$ that is continuous at θ_0 and $\sup_{\theta \in \Theta} \|\nabla_{\theta}\hat{g}_n(\theta) - G(\theta)\| \xrightarrow{P} 0$, (e) for $G = G(\theta_0)$, $G'WG$ is nonsingular. Then,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, (G'WG)^{-1}G'W\Omega WG'(G'WG)^{-1}).$$

The loss $\hat{L}_n(\theta)$ in Theorem S3 corresponds to our $\hat{L}_n(A)$. Our $\hat{g}_n(A)$ corresponds to their $\hat{g}_n(\theta)$. We have $\hat{A}_{W_n} \xrightarrow{P} \tilde{A}_0 = QA_0$ for some $Q \in \text{SP}(d)$ by Proposition 6.2, and the conditions on the weighting matrix are satisfied by (ii). Condition (a) of Theorem S3 is satisfied by assumption (v). For (b) note that $\hat{g}_n(A)$ is a polynomial map in A and hence smooth. For (c), by Lemma S7, $\sqrt{n} \text{vec}(\hat{m}_n(\tilde{A}_0) - m(\tilde{A}_0))$ weakly converges to $N(0, \Sigma_h^{2,r})$, where $h = \mu$ or $h = k$ pending whether moments or k-statistics are used to compute $\hat{m}_n(A)$. The variance matrices are defined in (S20) or (S24). However, $\hat{g}_n(\tilde{A}_0)$ is simply a projection of $(\hat{m}_n(\tilde{A}_0) - m(\tilde{A}_0))$ onto the coordinates of \mathcal{V}^\perp . Therefore, it also weakly converges to $N(0, \Sigma)$, where

$$(S27) \quad \Sigma = D_{\mathcal{I}}^{2,r} \Sigma_h^{2,r} D_{\mathcal{I}}^{2,r'}$$

with $D_{\mathcal{I}}^{2,r}$ being a selection matrix that selects the corresponding to the unique entries in $S^r(\mathbb{R}^d) \oplus \mathcal{V}^\perp$. Note that the specific form of Σ depends on whether moment or cumulant restrictions are used, i.e. $h = \mu, \kappa$. Here we suppress this dependence in the notation, but in Appendix ?? where we discuss the estimation of Σ we make it explicit.

We now show that (d) holds. The derivative of the map $g_{S,T}(A)$ in (16) is a linear mapping from $\mathbb{R}^{d \times d}$ to \mathbb{R}^{d_g} . It is obtained as a composition of the derivative of $m_{S,T}(A)$ given by the vectorized version of $(K_{S,A}(V), K_{T,A}(V))$,

with each component defined in (S1), and the projection $\pi_{\mathcal{V}}$. Thus, the derivative is given by mapping $V \in \mathbb{R}^{d \times d}$ to the vector

$$\text{vec}\left((V, A) \bullet S + (A, V) \bullet S, \pi_{\mathcal{V}}((V, A, \dots, A) \bullet T + \dots + (A, \dots, A, V) \bullet T)\right).$$

The Jacobian matrix $G_{S,T}(A)$ representing this derivative has d^2 columns and the column corresponding to variable A_{ij} is obtained simply by evaluating the derivative at the unit matrix $E_{ij} \in \mathbb{R}^{d \times d}$. In symbols, this column is given by stacking the vector $(E_{ij} \otimes A + A \otimes E_{ij})\text{vec}(S)$ over the vector

$$(S28) \quad ((E_{ij} \otimes A \otimes \dots \otimes A) + \dots + (A \otimes \dots \otimes A \otimes E_{ij})) \cdot \text{vec}(T),$$

and then selecting only the entries corresponding to the 2-tuples $i \leq j$ and r -tuples in \mathcal{I} .

Denote the Jacobian $G_{S,T}$ by $G(A)$ if $S = \mu_2(Y)$, $T = \mu_r(Y)$ and by $\hat{G}(A)$ if $S = \hat{\mu}_2$, $T = \hat{\mu}_r$ (or $S = \kappa_2(Y)$, $T = \kappa_r(Y)$ and $S = \mathbf{k}_2$, $T = \mathbf{k}_r$). The columns of $\hat{G}(A) - G(A)$ are like explained in (S28) with $S = \hat{\mu}_2 - \mu_2(Y)$ and $T = \hat{\mu}_r - \mu_r(Y)$ (or $S = \mathbf{k}_2 - \kappa_2(Y)$ and $T = \mathbf{k}_r - \kappa_r(Y)$). Since $\|S\| \xrightarrow{p} 0$ and $\|T\| \xrightarrow{p} 0$ by Lemma S7 part 1, and because A is fixed, the norm of each column converges to zero. In consequence, for each A , $\|\hat{G}(A) - G(A)\| \xrightarrow{p} 0$. Since \mathcal{A} is compact and $\hat{G}(A) - G(A)$ is smooth, we conclude

$$(S29) \quad \sup_{A \in \mathcal{A}} \|\hat{G}(A) - G(A)\| \xrightarrow{p} 0.$$

This establishes part (d). To establish part (e) note that W is positive definite and the Jacobian $G(QA_0)$ has full column rank by Lemma S4 below.

Lemma S4. *If \mathcal{V} assures identifiability up to a sign permutation matrix, then the matrix $G(QA_0)$ has full column rank for each $Q \in \text{SP}(d)$.*

Proof. It is enough to show that the derivative of $g(A)$ at QA_0 has trivial kernel. We first analyze the $S^2(\mathbb{R}^d)$ -part of the derivative noting that $\mu_2(Y) = \kappa_2(Y)$ as $\mathbb{E}Y = 0$. Suppose $(QA_0) \bullet \kappa_2(Y) = I_d$ and so the condition $(V, QA_0) \bullet \kappa_2(Y) + (QA_0, V) \bullet \kappa_2(Y) = 0$ is equivalent to

$$(A_0^{-1}Q'V, I_d) \bullet I_d + (I_d, A_0^{-1}Q'V) \bullet I_d = 0.$$

Using the derivative $K_{S,A}$ notation given in (S1), we write this last condition as $K_{I_d, I_d}(A_0^{-1}Q'V) = 0$. Similarly, the \mathcal{V}^\perp -part implies that $K_{T, I_d}(A_0^{-1}Q'V) = 0$ with $T = \kappa_r(Y)$. This implies that $A_0^{-1}Q'V = 0$ by Lemma S5 and the fact that I_d is an isolated point of \mathcal{G}_T . We conclude that V must be zero. \square

Having verified all conditions of S3 we can apply the theorem to prove the first display in Proposition 6.3. The second display follows as a special case when taking $W_n = \hat{\Sigma}_n^{-1}$, noting that $\hat{\Sigma}_n^{-1} \rightarrow \Sigma^{-1}$, and replacing W by Σ^{-1} in the first display.

S7.4. Proof of Proposition S1. Let $\tilde{A}_0 = QA_0$. Noting that $\hat{g}_n(\hat{A}_{\hat{\Sigma}_n^{-1}})$ minimizes $\|\cdot\|_{W_n}^2$ when taking $W_n = \hat{\Sigma}_n^{-1}$, we get that $\hat{g}_n(\hat{A}_{\hat{\Sigma}_n^{-1}}) = 0$. Using Taylor's theorem we get that

$$0 = \hat{\Sigma}_n^{-1/2} \sqrt{n} \hat{g}_n(\hat{A}_{\hat{\Sigma}_n^{-1}}) = \hat{\Sigma}_n^{-1/2} \sqrt{n} \hat{g}_n(\tilde{A}_0) + \hat{\Sigma}_n^{-1/2} \hat{G}(\bar{A}) \sqrt{n} \text{vec}(\hat{A}_{\hat{\Sigma}_n^{-1}} - \tilde{A}_0),$$

where \bar{A} lies on the segment between \tilde{A}_0 and $\hat{A}_{\hat{\Sigma}_n^{-1}}$. Pre-multiplying by $\hat{G}(\bar{A})' \hat{\Sigma}_n^{-1/2}$ and rearranging gives

$$\sqrt{n} \text{vec}(\hat{A}_{\hat{\Sigma}_n^{-1}} - \tilde{A}_0) = -[\hat{G}(\bar{A})' \hat{\Sigma}_n^{-1} \hat{G}(\bar{A})]^{-1} \hat{G}(\bar{A})' \hat{\Sigma}_n^{-1} \sqrt{n} \hat{g}_n(\tilde{A}_0).$$

Substituting $\sqrt{n} \text{vec}(\hat{A}_{\hat{\Sigma}_n^{-1}} - \tilde{A}_0)$ back into the expansion above gives

$$\hat{\Sigma}_n^{-1/2} \sqrt{n} \hat{g}_n(\hat{A}_{\hat{\Sigma}_n^{-1}}) = \hat{N} \hat{\Sigma}_n^{-1/2} \sqrt{n} \hat{g}_n(\tilde{A}_0)$$

where

$$\hat{N} = I_{d_g} - \hat{\Sigma}_n^{-1/2} \hat{G}(\bar{A}) [\hat{G}(\bar{A})' \hat{\Sigma}_n^{-1} \hat{G}(\bar{A})]^{-1} \hat{G}(\bar{A})' \hat{\Sigma}_n^{-1/2}.$$

By the discussion preceding (S27), we have $\Sigma^{-1/2} \sqrt{n} \hat{g}_n(\tilde{A}_0) \xrightarrow{d} Z \sim N(0, I_{d_g})$. Note that this random variable differs from $\hat{\Sigma}_n^{-1/2} \sqrt{n} \hat{g}_n(\tilde{A}_0) \xrightarrow{d} Z \sim N(0, I_{d_g})$ only by something that converges to zero in probability, as $\hat{\Sigma}_n \xrightarrow{p} \Sigma$. By Slutsky's lemma we have $\hat{\Sigma}_n^{-1/2} \sqrt{n} \hat{g}_n(\tilde{A}_0) \xrightarrow{d} Z \sim N(0, I_{d_g})$, and from Proposition 6.2, equation (S29) and $\hat{\Sigma}_n \xrightarrow{p} \Sigma$ and the continuous mapping theorem, we get

$$(S30) \quad \hat{N} \xrightarrow{p} N = I_{d_g} - \Sigma^{-1/2} G(\tilde{A}_0) [G(\tilde{A}_0)' \Sigma^{-1} G(\tilde{A}_0)]^{-1} G(\tilde{A}_0)' \Sigma^{-1/2}.$$

We note that N is a projection matrix of rank $d_g - d^2$. Combining we get

$$\begin{aligned} \hat{L}_{\hat{\Sigma}_n^{-1}}(\hat{A}_{\hat{\Sigma}_n^{-1}}) &= \left(\hat{\Sigma}_n^{-1/2} \sqrt{n} \hat{g}_n(\hat{A}_{\hat{\Sigma}_n^{-1}}) \right)' \left(\hat{\Sigma}_n^{-1/2} \sqrt{n} \hat{g}_n(\hat{A}_{\hat{\Sigma}_n^{-1}}) \right) \\ &\xrightarrow{d} Z' N Z \sim \chi^2(d_g - d^2), \end{aligned}$$

where the last step follows from Rao [1973, page 186].

S7.5. Proof of Proposition S2. From the proof of Proposition S1 we have

$$\hat{\Sigma}_n^{-1/2} \sqrt{n} \hat{g}_n(\hat{A}_{\hat{\Sigma}_n^{-1}}) = N \Sigma^{-1/2} \sqrt{n} \hat{g}_n(\tilde{A}_0) + o_p(1),$$

where N is the projection matrix defined in (S30). Let $\hat{g}_{1,n}$, G_1 , N_1 be the equivalent quantities to \hat{g}_n , G , N just computed for the smaller set of identifying restrictions. Using similar arguments we get

$$\begin{aligned} \hat{\Sigma}_{11}^{-1/2} \sqrt{n} \hat{g}_{1,n}(\hat{A}_{\hat{\Sigma}_{11}^{-1}}) &= N_1 \Sigma_{11}^{-1/2} \sqrt{n} \hat{g}_{1,n}(\tilde{A}_0) + o_p(1) \\ &= N_1 \Sigma_{11}^{-1/2} [I_{d_{g_1}} : 0_{d_{g_1} \times d_g}] \Sigma^{1/2} \Sigma^{-1/2} \sqrt{n} \hat{g}_n(\tilde{A}_0) \\ &\quad + o_p(1). \end{aligned}$$

Define $\Xi = \Sigma_{11}^{-1/2} [I_{d_{m_1}} : 0_{d_{g_1} \times d_g}] \Sigma^{1/2}$ and $J = N_1 \Xi$. Note that N is idempotent and set $B \equiv J' J = \Xi' N_1 \Xi$. We show that (i) $N - B$ is idempotent

and (ii) $N - B$ has rank $d_g - d_{g_1}$. First, letting $N = I_{d_g} - P$ with $P = \Sigma^{-1/2}G(\tilde{A}_0)[G(\tilde{A}_0)'\Sigma^{-1}G(\tilde{A}_0)]^{-1}G(\tilde{A}_0)'\Sigma^{-1/2}$, we have

$$\begin{aligned} BN &= B - BP(P'P)^{-1}P' \\ &= B - \Xi'N_1\Xi P(P'P)^{-1}P', \end{aligned}$$

and $N_1\Xi P = N_1P_1 = 0$, such that $BN = B$. Using similar step we find that $NB = B$. Finally, consider BB for which we have

$$\begin{aligned} BB &= \Xi'N_1\Xi\Xi'N_1\Xi \\ &= \Xi'N_1\Sigma_{11}^{-1/2}\Sigma_{11}\Sigma_{11}^{-1/2}N_1\Xi \\ &= \Xi'N_1\Xi = B \end{aligned}$$

Combining we get that $(N - B)(N - B) = N - B$. For (ii) note that since $N - B$ is idempotent we have $\text{rank}(N - B) = \text{Tr}(N - B) = d_g - d_{g_1}$. To complete the proof note that

$$\begin{aligned} C_n &= \sqrt{n}\hat{g}_n(\tilde{A}_0)'\hat{\Sigma}_n^{-1/2'}[N - B]\hat{\Sigma}_n^{-1/2}\sqrt{n}\hat{g}_n(\tilde{A}_0) + o_p(1) \\ &\xrightarrow{d} Z'[N - B]Z \sim \chi^2(d_g - d_{g_1}) . \end{aligned}$$

REFERENCES

- Francis R. Bach and Michael I. Jordan. Beyond independent components: Trees and clusters. *Journal of Machine Learning Research*, 4:1205–1233, 2003.
- David R Brillinger. The calculation of cumulants via conditioning. *Annals of the Institute of Statistical Mathematics*, 21(1):215–218, 1969.
- 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.
- Aiyu Chen and Peter J. Bickel. Efficient independent component analysis. *The Annals of Statistics*, 34(6):2825 – 2855, 2006.
- David Cox, John Little, and Donal OShea. *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer Science & Business Media, 2013.
- Elvira Di Nardo, Giuseppe Guarino, and Domenico Senato. A new method for fast computing unbiased estimators of cumulants. *Statistics and Computing*, 19(2):155–165, 2009.
- Morris L. Eaton. *Multivariate statistics*, volume 53 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. 2007. A vector space approach, Reprint of the 1983 original [MR0716321].
- Ronald Aylmer Fisher. Moments and product moments of sampling distributions. *Proceedings of the London Mathematical Society*, 2(1):199–238, 1930.
- Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.
- 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.
- A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.
- Aapo Hyvärinen, Patrik O Hoyer, and Mika Inki. Topographic independent component analysis. *Neural computation*, 13(7):1527–1558, 2001.
- 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. doi: 10.1007/978-3-642-15995-4_29.
- Sreenivasa Rao Jammalamadaka, Emanuele Taufer, and György H. Terdik. Asymptotic theory for statistics based on cumulant vectors with applications. *Scandinavian Journal of Statistics*, 48(2):708–728, 2021.

- 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.
- Whitney K. Newey and Daniel McFadden. Chapter 36 large sample estimation and hypothesis testing. volume 4 of *Handbook of Econometrics*, pages 2111–2245. Elsevier, 1994.
- C. Radhakrishna Rao. *Linear Statistical Inference and its Applications: Second Editon*. John Wiley & Sons, Inc., 1973.
- Richard J. Samworth and Ming Yuan. Independent component analysis via nonparametric maximum likelihood estimation. *The Annals of Statistics*, 40(6):2973 – 3002, 2012.
- Xiaofeng Shao and Jingsi Zhang. Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, 109(507):1302–1318, 2014.
- 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.
- Terry P. Speed. Cumulants and partition lattices 1. *Australian Journal of Statistics*, 25(2):378–388, 1983.
- Terry P. Speed. Cumulants and partition lattices ii: Generalised k-statistics. *Journal of the Australian Mathematical Society*, 40(1):34–53, 1986.
- Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769 – 2794, 2007.
- 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`