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 the moments or cumulants of ε. We document minimal conditions for identifiability and propose efficient estimation methods based on the new identification results. In situations where independence cannot be assumed the efficiency gains can be significant relative to methods that rely on independence. The proof strategy employed highlights new geometric and combinatorial tools that can be adopted to study identifiability via higher order restrictions in linear systems.

1. INTRODUCTION

Consider the linear system

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

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

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 principal as the interest is explicitly in recovering the independent components, see for instance the cocktail party problem described in Hyvärinen et al. [2001, p. 148]. However, in other applications, where the interest is solely in recovering A, the independence assumption is not a crucial starting point and can in fact be restrictive as the distribution of Y may not admit a linear transformation that leads to independent components [e.g. Matteson and Tsay, 2017, Kilian and Lütkepohl, 2017, Montiel Olea

Date: July 29, 2022.

et al., 2022]. Prominent examples include (trans)elliptical distributions and mixtures-of-Gaussians where generally no linear mapping to independent non-Gaussian components exists; see Section 2 for precise statements.

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

We systematically study our question by considering different restrictions on the higher order moments or cumulants of ε. We focus on cases where a subset of entries of a given rth order moment/cumulant tensor are set to zero, for some $r > 2$. Although there are alternative types of restrictions that can be considered, zero restrictions on moments are attractive as they often arise naturally from subject specific knowledge, whereas zero restrictions on cumulants can be adopted to characterize various relaxed notions of independence such as mean independence for instance.[1]

We distinguish between two types of identification results: (i) restrictions that imply that the identified set is exactly the set of signed permutation matrices and (ii) restrictions that imply that the identified set is finite. A general important finding is that there exists a substantial gap between (i) and (ii) in the sense that identified finite sets are often substantially larger when compared to the set of signed permutations.

We provide two classes of higher order restrictions that identify the set of signed permutation matrices. 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.[2] We show that, without imposing the independence assumption, if we set the off-diagonal elements of *any* rth order moment or cumulant tensor to zero we obtain sufficient identifying restrictions to pin down Q up to sign and permutation. We point out that for $r = 3, 4$ similar results are shown for moment restrictions in Guay [2021] and Velasco [2022] using a different proof strategy.[3]

Second, while off-diagonal zero restrictions are commonly adopted, they cannot always be used when the components of ε are not independent. For

[1] Section 4 explicitly relates zero restrictions on higher order cumulants to mean independence and additionally shows how invariance properties of the distribution of ε can also motivate zero moment/cumulant restrictions.

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

[3] Specifically, they show that the identified set is equal to the set of signed permutations using direct calculations for $r = 3, 4$. As we show below such calculations do not scale easy for higher order restrictions, nor non-diagonal restrictions.

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. 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.

Overall, diagonal and reflectionally invariant restrictions are most relevant for practical purposes, especially for low order tensors $r = 3, 4$, as efficient estimation methods can be easily implemented based on such identifying assumptions.

Next, we turn to exploring the identification problem in its full generality. We provide the minimal zero restrictions on the higher order moments (or cumulants) that ensure that the identified set is finite. It turns out that these restrictions are easy to understand and interpret, but the resulting identified set can be difficult. We illustrate this "identification gap", i.e. the gap between finite and sign-permutation sets, using simple examples and discuss possible additional restrictions that may close the gap in general settings.

Based on our identification results 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 the JADE algorithm [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. Moreover, we show that a particular subset of estimators in this class is efficient for model (1) when the only identifying restrictions are the zero cumulant/moment restrictions. Finally, we provide several hypothesis tests for the ex-post verification of the validity of the zero restrictions. These tests are useful to (a) validate the overall model specification and (b) evaluate specific subsets of zero restrictions.

The ICA literature is large and for a comprehensive review we refer to Hyvärinen et al. [2001] and more recently Comon and Jutten [2010]. Overall, our main contribution is to study the identification problem for model (1) when the components of ε are not independent.

The starting observation — independent components may not exists — is not new to this paper. In fact, such concerns were common in the early ICA literature, see Comon and Jutten [2010, Chapter 1] for an illuminating discussion, and they motivated explicit tests for the existence of independent

components [e.g. Oja et al., 2016, Matteson and Tsay, 2017, Davis and Ng, 2022].

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 [2002a], 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.

Finally, we note that our approach is different from methods that introduce an explicit alternative dependence structure on ε and aim to recover A with respect to this structure, for instance Cardoso [1998] and Hyvärinen and Hoyer [2000] group the components of ε into independent groups and Bach and Jordan [2002b] impose a tree-structured graphical model on ε. In contrast, our approach does not pre-specify a particular structure on ε, but rather investigates which types of higher order restrictions, can yield identification.

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 Sections 5 and 6. Inference is discussed in Section 7 followed by some numerical results in Section 8.

2. Motivation: independent components may not exist

Independent components analysis assumes that there exists a linear transformation of Y that results in a set of independent components ε. For Gaussian Y this is indeed true, but for many non-Gaussian distributions it is not. We briefly mention a few basic examples.

2.1. **Elliptical and transelliptical distributions.**

A particularly broad class of distributions where the existence of independent components largely fails is the elliptical class; see Kelker [1970] for a detailed discussion.

Definition 2.1. A random vector $X \in \mathbb{R}^d$ has an elliptical distribution if there exists $\mu \in \mathbb{R}^d$ and a positive semi-definite matrix Σ such that the characteristic function of X is of the form $\boldsymbol{t} \mapsto \phi(\boldsymbol{t}'\Sigma\boldsymbol{t})\exp(i\mu'\boldsymbol{t})$ for some $\phi : [0,\infty) \to \mathbb{R}$. A spherical distribution is an elliptical distribution with $\mu = 0$ and $\Sigma = I_d$.

Important special cases include the multivariate normal, Laplace and t-distributions. The following result implies that linear transformations that lead to independent components generally do not exist in the elliptical world.

Proposition 2.2. *A linear transformation of an elliptical random vector Y has independent components if and only if Y is Gaussian.*

Proof. The family of elliptical distributions is closed under taking linear combinations. So if Y is elliptical then so is AY. By Lemma 5 of Kelker [1970] AY can have independent components if and only if it is Gaussian. But then Y must be Gaussian. □

By Proposition 2.2, if Y is elliptical (but not Gaussian) and $AY = \varepsilon$ for some ε with $\mathbb{E}\varepsilon = 0$ and $\mathrm{var}(\varepsilon) = I_d$, then there exist no independent components and the classical ICA identification result of Comon [1994] cannot be used to establish identification of A.

That said, for elliptical distributions A can never be recovered beyond the set $\Omega = \{QA : Q \in \mathrm{O}(d)\}$. This follows directly because the distribution of a spherical distribution is invariant under the orthogonal transformations. In our results in Sections 5 and 6 this is reflected by the fact that moment/cumulant tensors of spherically distributed random variables do not satisfy the required genericity conditions.

In contrast, for any distribution of Y which is not *exactly* invariant to orthogonal transformations we can envision the existence of moment/cumulant restrictions that allow to improve the identification of A. As an example, consider the following definition for transelliptical distributions (elliptical copula models) [e.g. Frahm et al., 2003, Liu et al., 2012].

Definition 2.3. A random vector $X \in \mathbb{R}^d$ is transelliptical if there exist some univariate monotone transformations $f_1, \ldots, f_d$ such that the vector $f(X) = (f_1(X_1), \ldots, f_d(X_d))$ is elliptical. A nonparanormal distribution is a special case of the transelliptical distribution where $f(X)$ is Gaussian.

Since elliptical $f(Y)$ cannot have independent components unless $f(Y)$ is Gaussian, also transelliptical Y cannot have independent components unless it is nonparanormal. It follows that classical ICA identification result can only be used when Y is paranormal. In all other cases no independent components exists.

However, for general choices of transformations $f_1, \ldots, f_d$, the vector Y is not elliptical and moment/cumulant restrictions can be sought that allow to shrink the set Ω. As an illustrative example we will show in Section 5.2 that in the case when the transformations $f_1, \ldots, f_d$ are antisymmetric, the

cumulants of ε have a particular reflectionally invariant structure that can be exploited to identify A up to sign and permutation. Similar calculations can be conducted for any other desired deviation from the elliptical family.

2.2. **Gaussian mixture.** Consider the following mixture of two zero-mean Gaussians:

$$X \sim \begin{cases} \mathcal{N}(0, \Sigma_1) & w.p. \quad 1-\gamma \\ \mathcal{N}(0, \Sigma_2) & w.p. \quad \gamma \end{cases}.$$

Linear transformations of X, say BX for an invertible $B \in \mathbb{R}^{d\times d}$, preserve the mixture distribution: BX is a mixture of zero-mean Gaussians with the same mixture parameter γ and the covariance matrices transformed accordingly.

The following result highlights the special cases under which two mixtures of Gaussian distributions admit independent components.

Proposition 2.4. *Suppose that X has a distribution which is a mixture of two zero mean Gaussian distributions with covariances Σ_1 and Σ_2. If the components of X are independent then* at most one *of the components of X is non-Gaussian. If exactly one is non-Gaussian then both Σ_1 and Σ_2 must be diagonal and differ in at most one (diagonal) entry.*

Proof. The proof is provided in Appendix B.2. □

Recall that in the classical ICA setting, the components of ε must be independent and at most one of them is Gaussian. If ε is assumed to have the mixture distribution we immediately see that standard ICA methods cannot be used for recovering A in (1).

Corollary 2.5. *For $d \geq 3$, we have that a linear transformation of a random vector Y, which follows a mixture of two zero mean Gaussian distribution, cannot satisfy the standard non-Gaussianity assumption for identification in ICA, i.e. the components of the linear transformation cannot be both (i) independent and (ii) at least $d-1$ have a non-Gaussian distribution.*

These examples motivate our study that aims to relax the identification conditions such that we can identify A for more kinds of distributions for ε.

3. Basic tensor notation

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

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

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

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

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

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

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

$$h_r(AX) = A \bullet h_r(X), \tag{2}$$

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

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

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

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

$$[(A,\ldots,Z)\cdot T]_{i_1\cdots i_r} \;=\; \sum_{j_1=1}^{n_1} \cdots \sum_{j_r=1}^{n_r} A_{i_1 j_1}\cdots Z_{i_r j_r} T_{j_1\cdots j_r}. \tag{3}$$

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

Remark 3.1. The multilinearity property (2) 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 (2) and so the definition of $h_r(X)$ can be extended beyond moments and cumulants if needed.

In this paper we only consider moments and cumulants: moments because often subject specific knowledge regarding appropriate moment restrictions is available and cumulants because they can characterize independence and various relaxed notions of independence. For instance, the following well-known characterization of independence is of importance in our work.

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

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

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

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

4. Identification with zero constraints

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

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

In this section we show how to exploit additional structure in some $h_r(X)$ to further shrink Ω.

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 (2) we have

$$h_r(AY) = A \bullet h_r(Y) = T \;, \tag{4}$$

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

$$(QA) \bullet h_r(Y) \;\in\; \mathcal{V} \;. \tag{5}$$

However,

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

and so (4) and (5) hold together if and only if $Q \bullet T \in \mathcal{V}$. For $T \in \mathcal{V}$, we define

$$\mathcal{G}_T(\mathcal{V}) \;:=\; \{Q \in \mathrm{O}(d) : Q \bullet T \in \mathcal{V}\} \;, \tag{6}$$

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

We summarize the general identification problem as follows.

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

$$\Omega_0 \;=\; \{QA : Q \in \mathcal{G}_T(\mathcal{V})\}. \tag{7}$$

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

Clearly, there exists a plethora of restrictions on the higher order moment or cumulants that can be considered. For instance, the ICA assumption imposes that $\kappa_r(\varepsilon) = \mathrm{cum}_r(\varepsilon)$ has zero off-diagonal elements for all r (i.e. Proposition 3.2). We will relax this assumption in two ways. First, from the definition of $\mathcal{V}$ it follows that we will consider only restrictions on a single moment or cumulant tensor $T = h_r(X)$ and second we will explore which off-diagonal elements can be made non-zero. In other words, within the class of zero restrictions we look for minimal ones.

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

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

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

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

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

In a matrix form this can be written as

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

Since Q is orthogonal, each of the two diagonal matrices above is either identically zero or it is invertible. If it is identically zero then Q must be a sign permutation matrix and the equation clearly holds. If they are both invertible we immediately see that the equation cannot hold unless

$T_{111} = T_{222} = 0$, in which case T is the zero tensor showing that for every nonzero $T \in \mathcal{V}$ we have that $\mathcal{G}_T(\mathcal{V}) = \mathrm{SP}(2)$.

The example clarifies our notation and illustrates how higher order moment or cumulant restrictions can be used for identification. Unfortunately, the direct arguments that we used 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 arises with 4th order moment restrictions arises when $\mathbb{E}\varepsilon = 0$, $\mathrm{var}(\varepsilon) = I_d$ and $T_{iijj} = \mathbb{E}\varepsilon_i^2\varepsilon_j^2 = 1$ for $i \neq j$.

4.2. **Deriving zero restrictions.** Zero restrictions on higher order moments often arise naturally from subject specific knowledge. A prominent example can be found in economics where canonical models allow to interpret the errors $\varepsilon = (\varepsilon_1, \varepsilon_2)$ as supply and demand shocks implying that $\mathbb{E}\varepsilon_1\varepsilon_2^l = 0$ for some $l \geq 1$, see Bekaert et al. [2021] for an example where such restrictions are exploited for $l = 1, 2, 3$. For zero cumulant restrictions there exists often less subject specific knowledge. Instead such restrictions arise naturally under alternative relaxations of independence.

Mean independence. We say that X_i is mean independent of X_j if $\mathbb{E}(X_i|X_j) = \mathbb{E}(X_i)$. Mean independence is strictly stronger than uncorrelatedness yet strictly weaker than independence. The relationship to zero cumulants can be formalized as follows.

Proposition 4.4. *If X_i is mean independent of X_j then $[\kappa_r(X)]_{ij\cdots j} = 0$ for every $r \geq 2$.*

Proof. By semi-invariance of cumulants of order $r \geq 2$, we can assume that $\mathbb{E}X_i = 0$. For every $l \in \mathbb{N}$ we have

$$\mathbb{E}(X_i X_j^l) = \mathbb{E}(X_j^l \mathbb{E}(X_i|X_j)) = \mathbb{E}(X_i)\mathbb{E}(X_j^l) = 0.$$

We can now use the formula (23) for cumulants in terms of moments, which we give in the appendix, to conclude the proof. □

There is an obvious generalization of this result that admits essentially the same proof.

Proposition 4.5. *If X_i is mean independent of a subvector X_B of X then $[\kappa_r(X)]_{ij_2\cdots j_r} = 0$ for every r and any collection $j_2, \ldots, j_r$ of elements in B.*

These results can be used directly to formulate zero cumulant restrictions for model (1) when, for instance, ε_i is known to be mean independent of some subvector ε_B.

Remark 4.6. Note that X_i is mean independent of X_B if and only if

$$\operatorname{cov}(X_i, f(X_B)) = 0$$

for any function $f : \mathbb{R}^{|B|} \to \mathbb{R}$ for which this covariance exists. As a further relaxation of mean independence we could require that $\operatorname{cov}(X_i, f(X_B)) = 0$ only for all polynomials f in x_B of order at most $r-1$. Using the same proof as in Proposition 4.4, we conclude that $[\kappa_r(X)]_{ij\cdots j} = 0$ (and similar restrictions may not hold for higher orders).

Invariance. Another natural way how zero moment/cumulant restrictions appear is through invariance properties on the underlying distribution. Suppose that the distribution of X is the same as the distribution of DX for *every* diagonal matrix D with $D_{ii} = \pm 1$ for all $i = 1, \ldots, d$ (e.g. when X has spherical distribution). In this case, by multilinearity of cumulants,

$$[h_r(DX)]_{i_1\cdots i_r} \;=\; D_{i_1 i_1} \cdots D_{i_r i_r} [h_r(X)]_{i_1\cdots i_r} \;.$$

Since D is arbitrary, $[\kappa_r(X)]_{i_1\cdots i_r}$ 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}. We treat this zero pattern more in detail in Section 5.2 and we show how it appears in (trans)elliptical distributions.

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.

5.1. Diagonal tensors.

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

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

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

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

$$f_T(x) = \sum_{\boldsymbol{i}} T_{\boldsymbol{i}} x_{i_1} \ldots x_{i_r} \;=\; \langle T, x^{\otimes r} \rangle, \tag{8}$$

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

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

Proof. The first claim follows because

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

The second claim is a direct check. □

This will be useful for deriving our first main result.

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

Proof. The left direction is clear. For the right direction, note that the tensor T is diagonal if and only if $\nabla^2 f_T(x)$ is 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

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

If each row of Q has exactly one non-zero entry then $Q \in \mathrm{SP}(d)$ and we are done. So suppose $Q_{ij}, Q_{ik} \neq 0$. Then, by the above equation

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

Equivalently, $\frac{\partial^2}{\partial x_j^2} f_T(x) = \frac{\partial^2}{\partial x_k^2} f_T(x)$, which simply states that

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

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

Remark 5.4. The genericity condition is not only sufficient but also necessary. Indeed, if, for example $T_{1 \cdots 1} = T_{2 \cdots 2} = 0$ then $\frac{\partial}{\partial x_1^2} f_T(x) = \frac{\partial^2}{\partial x_2^2} f_T(x) = 0$. Thus, $\nabla^2 f_{Q \bullet T}(x)$ is diagonal for any block matrix of the form

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

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

Combining Proposition 4.1 and Theorem 5.3 implies the following result.

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

5.2. **Reflectionally invariant tensors.** In some applications the assumption that T is diagonal may be unattractive. A leading example is found in economics and finance where errors often display excess kurtosis, but at the same time the variance of the errors has a common component, e.g. the variance underlying stock market returns co-moves across different assets. This implies that exploiting 4th order moments may provide identifying information, but entries of the form T_{iijj} cannot be restricted to zero (or some other constant) due to the common volatility structure [e.g. Montiel Olea et al., 2022].

More generally, suppose that the distribution of ε is symmetric in the sense that its distribution will not change if we switch signs of some of its components. In this case the odd-order cumulants must be necessarily zero so, although they are diagonal, they do not satisfy the genericity conditions in Theorem 5.5. Hence we only consider even order tensors, which are generically not diagonal in this class of distributions.

These observations motivate the following tensor restrictions.

Definition 5.6. A tensor $T \in S^r(\mathbb{R}^d)$ is called *reflectionally invariant* if the only potentially non-zero entries in T are the entries $T_{i_1 \cdots i_r}$ where each index appears in the sequence $(i_1, \ldots, i_r)$ even number of times. If r is odd, the only reflectionally invariant tensor is the zero tensor.

Specific sub-classes of symmetric distributions can be shown to imply reflectionally invariant tensors. For instance if ε is spherical, all even moment/cumulant tensors are reflectionally invariant. A more surprising result is obtained for certain transelliptical distributions .

Proposition 5.7. *Suppose that ε is a random vector such that there exist strictly monotone and* antisymmetric *transformations $f_1, \ldots, f_d$ such that $f(\varepsilon) = (f_1(\varepsilon_1), \ldots, f_d(\varepsilon_d))$ is spherical. Then every even moment/cumulant tensor of ε is reflectionally invariant.*

Proof. It is enough that the distribution of ε is equal to the distribution of $D\varepsilon$ for every sign matrix D. The distribution of ε is characterized by the fact that $f(\varepsilon)$ has the same distribution as a spherical variable Z. On the other hand, since $f_1, \ldots, f_d$ are antisymmetric, $f(D\varepsilon) = Df(\varepsilon)$. However, the distribution of $Df(\varepsilon)$ is the same as the distribution of $f(\varepsilon)$ as it is spherical. It follows that the distribution of $f(D\varepsilon)$ is the same as the distribution of $f(\varepsilon)$ and consequently the distribution of $D\varepsilon$ equals the distribution of ε. □

To prove that reflectionally invariant tensors can be used to identify A in (1), recall from (8) that any $T \in S^r(\mathbb{R}^d)$ has an associated homogeneous polynomial $f_T(x)$ of order r in $x = (x_1, \ldots, x_d)$. 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 $f_T(x) = g_T(x_1^2, \ldots, x_d^2)$. The polynomial g_T is a polynomial associated to the tensor $S \in S^l(\mathbb{R}^d)$ defined by $S_{i_1 \cdots i_l} = T_{i_1 i_1 \cdots i_l i_l}$. We have the following useful characterization of reflectionally invariant tensors.

Lemma 5.8. *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

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

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 $g_T(x_1^2, \ldots, x_d^2)$. This is equivalent with T being reflectionally invariant. □

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

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

Theorem 5.9. *Suppose that $T \in S^r(\mathbb{R}^d)$ for an even r is a reflectionally invariant tensor satisfying*

$$T_{+\cdots+ii} \neq T_{+\cdots+jj} \quad \text{for all } i \neq j. \tag{9}$$

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

Remark 5.10. We emphasize that the genericity condition in (9) simply states that T lies outside of $\binom{d}{2}$ explicit linear hyperplanes in $S^r(\mathbb{R}^d)$. None of these hyperplanes contains the linear space of reflectionally invariant tensors and so the underlying measure of reflectionally invariant tensors not satisfying (9) is zero.

Theorem 5.9 is proven using the following lemma.

Lemma 5.11. *Let r be even and suppose that $T \in S^r(\mathbb{R}^d)$ is reflectionally invariant tensor satisfying (9). Then $Q \bullet T = T$ for $Q \in \mathrm{O}(d)$ if and only if Q is a diagonal matrix.*

Proof. The left implication is clear because $f_T(x) = f_T(Dx) = f_{D \bullet T}(x)$ by Lemma 5.8. 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 (9), 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 (3) as

$$(Q, \ldots, Q, I_d, I_d) \cdot T \;=\; (I_d, \ldots, I_d, Q', Q') \cdot T. \tag{10}$$

We want to show that this equality implies that Q is a diagonal matrix. Let $\boldsymbol{i} = (i_1, \ldots, i_r)$ and consider all $(r+2)$-tuples $(\boldsymbol{i}, u, u)$ for some $u \in \{1, \ldots, d\}$. Writing (10) restricted to these indices gives

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

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

$$\sum_{j_1,\ldots,j_r} Q_{i_1 j_j} \cdots Q_{i_r j_r} S_{j_1 \cdots j_r} \;=\; \sum_{v} T_{i_1 \cdots i_r vv} \;=\; S_{i_1 \cdots i_r}.$$

Since this equation holds for every $\boldsymbol{i} = (i_1, \ldots, i_r)$, we conclude $Q \bullet S = S$, where $S = (S_{\boldsymbol{i}}) \in S^r(\mathbb{R}^d)$. Note however that S is a reflectionally invariant tensor. Indeed, if some index appears in $\boldsymbol{i}$ odd number of times then $S_{\boldsymbol{i}} = T_{\boldsymbol{i}uu} = 0$ as the same index appears in $(\boldsymbol{i}, u, u)$ odd number of times. Since T satisfies (9), S satisfies (9) too. Indeed,

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

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

Proof of Theorem 5.9. The left implication is clear. For the right implication, suppose $Q \in \mathrm{O}(d)$ is such that $Q \bullet T$ is reflectionally invariant. By Lemma 5.8, equivalently, $f_{Q \bullet T}(x) = f_{Q \bullet T}(Dx)$ for every diagonal $D \in \mathrm{O}(d)$, which gives $f_T(Q'x) = f_T(Q'Dx)$. This polynomial equation implies that

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

but since $Q'DQ \in O(d)$, Lemma 5.11 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. We conclude 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 $Q \in \mathrm{SP}(d)$. □

Combining Proposition 4.1 and Theorem 5.9 implies the following result.

Theorem 5.12. *Consider the model (1) with $\mathbb{E}\varepsilon = 0$, $\mathrm{cov}(\varepsilon) = I_d$ and suppose that for some even r the tensor $h_r(\varepsilon)$ is reflectionally invariant and it satisfies the genericity condition (9). Then A is identifiable up to permuting and swapping signs of its rows.*

5.3. **Generalizations.** Theorems 5.5 and 5.12 highlight key zero moment and cumulant patterns that can be used to identify A up to sign and permutation for model (1). Obviously, such restrictions are equally sufficient for identification in the class of linear simultaneous equations models $AY = BX + \varepsilon$ when X is exogenous, and various dynamic extensions of such models [e.g. Kilian and Lütkepohl, 2017].

That said it is also of interest to explore whether relaxing additional zero restrictions still leads to identification (up to sign and permutation), and which genericity conditions are required. Since $\dim(\mathrm{O}(d)) = \binom{d}{2}$, we need at least that many constraints to assure $\mathcal{G}_T$ is finite. However, as we show in Section 6, this may still not be enough to assure that $\mathcal{G}_T = \mathrm{SP}(d)$. Motivated by Proposition 4.4, we consider a special model with

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

Conjecture 5.13. If T is a generic tensor in $\mathcal{V}(\mathcal{I})$ then $Q \bullet T \in \mathcal{V}(\mathcal{I})$ if and only if $Q \in \mathrm{SP}(d)$.

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

Proposition 5.14. *Suppose that $T \in S^r(\mathbb{R}^2)$ satisfies $T_{12\cdots 2} = T_{1\cdots 12} = 0$ but is otherwise generic. Then $Q \bullet T \in \mathcal{V}(\mathcal{I})$ if and only if $Q \in \mathrm{SP}(2)$.*

We prove this result in Appendix B.3. The genericity conditions are again linear and can be recovered from the proof.

6. Local identification

The results in the previous section 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. In general however, such identification results may be hard to obtain and, as we illustrate below, for many zero patterns a generic $\mathcal{G}_T$ will be finite but much more complicated than the set of sign permutations.

This section gives the 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. In addition, we note that for some applications it may not be necessary to recover A up to sign and permutation only, but rather knowing that $\mathcal{G}_T$ is finite suffices as it ensures that, for instance, a given estimation method converges.

Let $\mathcal{V} \subset S^r(\mathbb{R}^d)$ be a set given by polynomial constraints. A subset $\mathcal{U} \subseteq \mathcal{V}$ is Zariski open in $\mathcal{V}$ if the complement $\mathcal{V} \setminus \mathcal{U}$ is given by some additional polynomial constraints. 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 (9) 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.

Definition 6.1. 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 $\mathcal{G}_T(\mathcal{U})$.

The following result establishes link between local identification and finiteness of $\mathcal{G}_T$.

Proposition 6.2. *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 $f_1(T) = \cdots = f_k(T) = 0$ are the polynomials describing $\mathcal{V}$ then the polynomials describing $\mathcal{G}_T(\mathcal{V})$ within $\mathrm{O}(d)$ are $f_1(Q \bullet T) = \cdots = f_k(Q \bullet T) = 0$. Similarly, if $\mathcal{V} \setminus \mathcal{U}$ is described within $\mathcal{V}$ by $g_1(T) = \cdots = g_l(T) = 0$. Then $\mathcal{G}_T(\mathcal{V}) \setminus \mathcal{G}_T(\mathcal{U})$ is described by $g_1(Q \bullet T) = \cdots = g_l(Q \bullet T) = 0$.

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 $Q \in \mathcal{G}_T(\mathcal{U})$ 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, $g_1(Q \bullet T) \neq 0, \ldots, g_l(Q \bullet T) \neq 0$. Thus, in any sufficiently small neighbourhood of Q there will be a point that lies in $\mathcal{G}_T(\mathcal{V})$ and $g_1, \ldots, g_l$ 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. □

Remark 6.3. The proof of Proposition 6.2 also shows that if $\mathcal{U}$ is a Zariski open subset of $\mathcal{V}$ 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 $\mathcal{G}_T(\mathcal{V})$.

Lemma 6.4. *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

$$K_{T,A}(V) \;=\; (V, A, \dots, A) \bullet T + \cdots + (A, \dots, A, V) \bullet T. \tag{11}$$

Moreover, if A is invertible, then

$$K_{T,A}(V) \;=\; K_{A \bullet T, I_d}(VA^{-1}). \tag{12}$$

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

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

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

For a given linear subspace $\mathcal{V} \subseteq S^r(\mathbb{R}^d)$, let $\pi_{\mathcal{V}} : S^r(\mathbb{R}^d) \to \mathcal{V}^\perp$ 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 $T_{\boldsymbol{i}}$ for $\boldsymbol{i} \in \mathcal{I}$.

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 (11).

Lemma 6.5. *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*

$$K_{I_d,Q}(V) = \pi_{\mathcal{V}}(K_{T,Q}(V)) = 0 \qquad \textit{implies } V = 0. \tag{13}$$

Proof. Since,

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

V is a direction in the tangent space to $\mathrm{O}(d)$ 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 $\mathrm{O}(d)$ at Q.

The proof of Proposition 6.2 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 $t \to 0$, 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. □

Remark 6.6. In the examples of Section 5, for $T \in \mathcal{U} \subseteq \mathcal{V}$, we always had $\mathcal{G}_T(\mathcal{U}) = \mathcal{G}_T(\mathcal{V}) = \mathrm{SP}(d)$. The proof of Proposition 6.2 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 $\mathcal{G}_T(\mathcal{V})$.

The main result of this section studies local identifiability with a model defined by the minimal number of $\binom{d}{2}$ constraints with

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

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

$$B^{(j)}_{kl} \;=\; [T_{klj\cdots j}]_{k,l<j} \;\in\; S^2(\mathbb{R}^{j-1}) \tag{14}$$

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

$$\det\left(T_{j\cdots j}I_{j-1} - (r-1)B^{(j)}\right) \neq 0 \quad \text{for all } j = 2,\ldots,d. \tag{15}$$

Theorem 6.7. *If $T \in \mathcal{U}^\circ$ then $|\mathcal{G}_T(\mathcal{U}^\circ)| < \infty$.*

Proof. By Proposition 6.2 it is enough to show that each point of $\mathcal{G}_T(\mathcal{U}^\circ)$ is isolated. By Lemma 6.5, 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 (12), $K_{I_d,Q}(V) \;=\; K_{I_d,I_d}(VQ')$. 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 (12), 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 $\boldsymbol{i} = (i,j,\ldots,j)$ with $i < j$, $(K_{S,I_d}(U))_{ij\cdots j} = 0$. More explicitly,

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

Let $u_j = (U_{j\,1},\ldots,U_{j\,j-1})$ for $j = 2,\ldots,d$. Let first $j = d$. Using the matrix $B^{(d)}$ defined in (14) the equation above gives

$$\left(S_{d\cdots d}I_{d-1} - (r-1)B^{(d)}\right)u_d \;=\; 0.$$

This has a unique solution $u_d = 0$ if and only if $\det(S_{d\cdots d}I_{d-1} - (r-1)B^{(d)}) \neq 0$, which holds by (15). We have shown that the last row of U is zero. Now suppose that we have established that the rows $j+1,\ldots,d$ of U are zero. If $j = 1$, we are done by the fact that U is antisymmetric. So assume $j \ge 2$. We will use the fact that $U_{jl} = 0$ if $l \ge j$. For every $i < j$

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

This again has a unique solution if and only if $\det(S_{j\cdots j}I_{j-1}-(r-1)B^{(j)}) \neq 0$, which holds by (15). Using a recursive argument, we conclude that $U = 0$. □

Example 6.8. 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

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

Going back to our original motivation, suppose ε is a two-dimensional mean-zero random vector with $\mathrm{var}(\varepsilon) = I_2$. 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 6.9. The set $\mathcal{G}_T(\mathcal{U}^\circ)$ is finite but, as illustrated by Example 6.8, it typically contains matrices that do not have an easy interpretation. In particular, if $d = 2$ then $\mathcal{V}^\circ$ is given by a single constraint $T_{12\cdots 2} = 0$. 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 6.10. *Consider the model* (1) *with* $\mathbb{E}\varepsilon = 0$, $\mathrm{var}(\varepsilon) = I_d$ *and suppose that either* $\mu_r(\varepsilon) \in \mathcal{U}^\circ$ *or* $\kappa_r(\varepsilon) \in \mathcal{U}^\circ$. *Then* A *is locally identifiable.*

7. 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 following two basic observations:

Observation 1: Suppose $h_1(\varepsilon) = 0$, $h_2(\varepsilon) = I_d$, $h_r(\varepsilon) \in \mathcal{V}$, and $\mathcal{V}$ is enough to identify A_0 in (1) up to sign permutations (as discussed in Section 5). Then $A = QA_0$ for $Q \in \mathrm{SP}(d)$ if and only if $A \bullet h_2(Y) = I_d$ and $A \bullet h_r(Y) \in \mathcal{V}$, where we recall that we may take $h_r(\varepsilon) = \mu_r(\varepsilon)$ or $h_r(\varepsilon) = \kappa_r(\varepsilon)$.

Observation 2: The last statement remains approximately true if we replace the moments μ or cumulants κ with consistent estimators. We can then estimate A by choosing it such that the distance between $(A \bullet h_2(Y), A \bullet h_r(Y))$ and $(I_d, \mathcal{V})$ is minimized.

Such minimum distance estimators are commonly adopted in the ICA literature using (a) Euclidean distance to measure distance and (b) diagonal

tensor restrictions [e.g. Hyvärinen et al., 2001, Chapter 11]. For instance, the JADE algorithm of Cardoso and Souloumiac [1993] solves a minimum distance problem that considers (after pre-whitening) cumulant restrictions on κ_4. In our approach we (a) measure the distance to $\mathcal{V}$ in a statistically meaningful way in order to get optimal efficiency of the associated estimator and (b) consider also non-diagonal tensor restrictions. We note that the minimum distance approach is most natural given our identification results, but other existing ICA methods could equally well be modified.

We formalize these ideas as follows. Let A_0 denote the true A. For any symmetric matrix $S \in S^2(\mathbb{R}^d)$ and a symmetric tensor $T \in S^r(\mathbb{R}^d)$ define $m_{S,T} : \mathbb{R}^{d\times d} \to S^2(\mathbb{R}^d) \oplus S^r(\mathbb{R}^d)$ to be

$$m_{S,T}(A) \;=\; (A \bullet S - I_d,\; A \bullet T)\;. \tag{16}$$

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 rth order k-statistic, denoted by k_r, which are computed from a given sample $\{Y_s\}_{s=1}^n$. The computation of the sample moments $\hat{\mu}_r$ requires no explanation and for k-statistics we refer to McCullagh [2018, Chapter 4] as well as Appendix A.3 where we provide explicit computational formulas. Moreover, in Appendix A.6 we provide asymptotic results for the sample statistics. 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.

Remark 7.1. The inclusion of the first term $A \bullet h_2(Y) - I_d$ in $m(A)$ is not necessary when the data are pre-whitened, but we will not assume this. Also, if $\mathbb{E}\varepsilon \neq 0$ we may include $h_1(AY)$ in $m(A)$.

Now fix $\mathcal{V} = \mathcal{V}(\mathcal{I})$ and recall that $\pi_{\mathcal{V}}$ was defined as the orthogonal projection from $S^r(\mathbb{R}^d)$ to $\mathcal{V}^\perp$. For a fixed $\mathcal{V} = \mathcal{V}(\mathcal{I})$ we also define

$$g_{S,T}(A) \;:=\; \mathrm{vec}_u(A \bullet S - I_d,\; \pi_{\mathcal{V}}(A \bullet T)) \;\in\; \mathbb{R}^{\binom{d+1}{2}+|\mathcal{I}|}, \tag{17}$$

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 as a vector. We have $g_{S,T}(A) = 0$ if and only if $A \bullet S = I_d$ and $A \bullet T \in \mathcal{V}$. Let

$$d_g \;=\; \binom{d+1}{2} + |\mathcal{I}|$$

denote the dimension of $S^2(\mathbb{R}^d) \oplus \mathcal{V}^\perp$. Under any set of identifying restrictions $|\mathcal{I}| \geq \binom{d}{2}$ and so we have $d_g \geq d^2$. As in the case of $m_{S,T}(A)$ we write $g(A)$ if $S = h_2(Y)$ and $T = h_r(Y)$, and $\hat{g}_n(A)$ if $S = \hat{h}_2$, $T = \hat{h}_r$.

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

$$L_W(A) \;=\; \|g(A)\|_W^2 \quad \text{and} \quad \hat{L}_W(A) \;=\; \|\hat{g}_n(A)\|_W^2, \tag{18}$$

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

Lemma 7.2. *Suppose that (1) holds with $h_2(\varepsilon) = I_d$ and $h_r(\varepsilon) \in \mathcal{V}$. If $\mathcal{G}_T(\mathcal{V}) = \mathrm{SP}(d)$ then $L_W(A) = 0$ if and only if $A = QA_0$ for $Q \in \mathrm{SP}(d)$.*

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

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

which implies that $Q \in \mathcal{G}_T(\mathcal{V}) = \mathrm{SP}(d)$. □

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

$$\widehat{A}_{W_n} := \arg\min_{A \in \mathcal{A}} \hat{L}_{W_n}(A), \tag{19}$$

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

For moment restrictions class of estimators (19) falls in the class of generalized moment estimators as proposed in Hansen [1982], see also Hall [2005]. For cumulant restrictions the class of estimators (19) include prominent existing cumulant tensor estimators for the ICA model as special cases, see Hyvärinen et al. [2001, Chapter 11] for examples.

7.1. **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 7.3 (Consistency)**.** *Suppose that $\{Y_s\}_{s=1}^n$ is i.i.d and (i) $g(A) = 0$ if and only if $A = QA_0$ for $Q \in \mathrm{SP}(d)$, (ii) $\mathcal{A} \subset GL(d)$ is compact and $QA_0 \in \mathcal{A}$ for some $Q \in SP(d)$ (iii) $W_n \xrightarrow{p} W$ and W is positive definite, (iv) $\mathbb{E}\|Y_s\|^r < \infty$. Then $\widehat{A}_{W_n} \xrightarrow{p} QA_0$ as $n \to \infty$ for some $Q \in \mathrm{SP}(d)$.*

The proof and all remaining proofs of this section are deferred to Appendix C. We note that being able to satisfy condition (i) is the main contribution of our paper (cf. Lemma 7.2). For instance, for $\mathcal{I}$ containing the off-diagonal tensor indices Theorem 5.5 shows that this condition holds. The other conditions are more standard. 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.

7.2. **Asymptotic normality.** The positive definite 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 more efficient estimates provided that sufficient moments of Y exist. Specifically, when we take W_n such that it is consistent for the inverse of

$$\Sigma = \lim_{n\to\infty} \operatorname{var}(\sqrt{n}\hat{g}_n(QA_0)) \tag{20}$$

we can ensure that the resulting estimate $\widehat{A}_{W_n}$ achieves minimal variance in the class of generalized cumulant estimators (19).

Let $G(A) \in \mathbb{R}^{d_g \times d^2}$ be the Jacobian matrix representing the derivative of the function $g : \mathbb{R}^{d\times d} \to \mathbb{R}^{d_g}$ defined in (17). Here, defining the Jacobian we think about g as a map from $\mathbb{R}^{d^2}$ vectorizing A.

Proposition 7.4 (Asymptotic normality)**.** *Suppose that the conditions of Proposition 7.3 hold, (v) $QA_0 \in \operatorname{int}(\mathcal{A})$ for some $Q \in SP(d)$, (vi) $\mathbb{E}\|Y_i\|^{2r} < \infty$, and denote by $G = G(QA_0)$. Then*

$$\sqrt{n}\operatorname{vec}[\widehat{A}_{W_n} - QA_0] \xrightarrow{d} N(0, (G'WG)^{-1}G'W\Sigma WG(G'WG)^{-1}) \tag{21}$$

for some $Q \in \mathrm{SP}(d)$, where Σ is given in (20). Moreover, for any $\widehat{\Sigma}_n \xrightarrow{p} \Sigma$ we have that

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

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

This result allows for an interesting comparison. For ICA models under full independence an efficient estimation method is developed in Chen and Bickel [2006]. When we relax the independence assumption, and instead only restrict higher order cumulant entries, the efficient estimator is given by $\widehat{A}_{\widehat{\Sigma}_n^{-1}}$. Here efficiency is understood in the sense that V is smaller when compared to the variance in (21) for any W. Chamberlain [1987] shows that for moment restrictions the estimator $\widehat{A}_{\widehat{\Sigma}_n^{-1}}$ attains the semi-parametric efficiency bound in the class of non-parametric models characterized by restrictions $T_{\boldsymbol{i}} = 0$ for $\boldsymbol{i} \in \mathcal{I}$.

Implementing this estimator can be done in different ways. Proposition 7.3 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 7.4). With this estimate we can compute $\widehat{A}_{\widehat{\Sigma}_n^{-1}}$ from (19). While this estimate is efficient, the procedure can obviously be iterated until convergence to avoid somewhat arbitrarily stopping at the first iteration, see Hansen and Lee [2021] for additional motivation for iterative moment estimators. Additionally, we may also consider $W_n = \widehat{\Sigma}_n(A)^{-1}$ as a weighting matrix, hence parametrizing the asymptotic variance estimate as a function of A, and minimize the objective function (19) 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 D.

7.3. **Testing over-identifying restrictions.** While zero restrictions on higher order moments or cumulants can be motivated from several angles (cf. the discussion in Section 4), 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 then 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 7.5. *If the conditions of Proposition 7.4 hold we have that as* $n \to \infty$

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

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

Proposition 7.6. *If the conditions of Proposition 7.4 hold we have that as* $n \to \infty$

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

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.

8. Numerical illustration

In this section we discuss the numerical implementation for estimators in the class of minimum distance estimators (19). We evaluate and compare the performance of several class members in different simulation designs.

8.1. **Diagonal tensors.** We start by investigating the usefulness of diagonal tensors $T^r = \kappa_r(\varepsilon)$, for orders $r = 3, 4$, for estimating A. We note that with normalized ε we have $\kappa_3(\varepsilon) = \mu_3(\varepsilon)$ and for $r = 4$ we found no systematic differences between using moments and cumulants. Therefore we only show the results for cumulant restrictions. To obtain samples $\{Y_s\}_{s=1}^n$

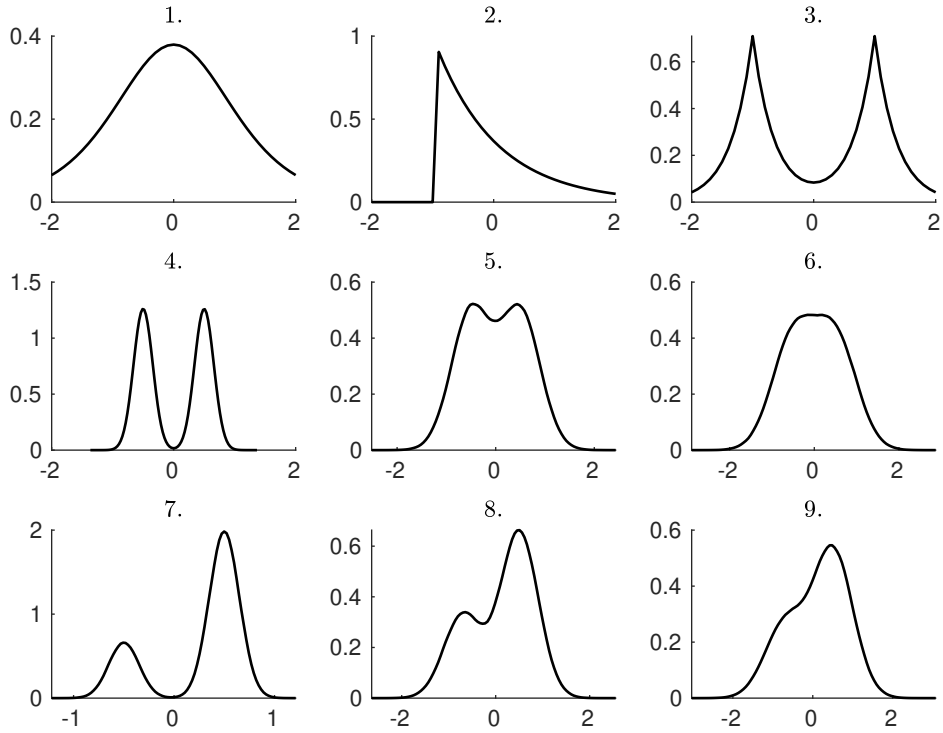

FIGURE 1. We show the univariate densities, as taken from Bach and Jordan [2002a], that were used for simulating errors for the diagonal cumulant tensor study.

from model (1) under different distributions we sample the errors independently from 9 different univariate distributions taken from Bach and Jordan [2002a] and reproduced in Figure 8.1. We take $d = 2, 3$ and a sample size of $n = 500$.

For each sample we estimate A by minimizing $\hat{L}_{W_n}(A)$ for $W_n = I_{d_g}$ and $W_n = \widehat{\Sigma}_n^{-1}$. For cumulant restrictions taking $W_n = I_{d_g}$ corresponds to the conventional choice in the ICA literature and for $r = 4$ we obtain the JADE algorithm. The asymptotically optimal choice $W_n = \widehat{\Sigma}_n^{-1}$ is motivated by Proposition 7.4. Computationally, we first estimate A based on $W_n = I_{d_g}$ with $d_g = \binom{d}{2} + \binom{d+r-1}{r} - d$, after which we compute $\widehat{\Sigma}_n$ following the discussion in Appendix D, and repeat the estimation procedure using $W_n = \widehat{\Sigma}_n^{-1}$. It is important to point out that for cumulant restrictions the $\mathtt{k}_r$ statistics that are needed to compute the entries of $\hat{g}_n(A)$ can be rapidly computed using the function `nPolyk` as provided in the `kStatistics` package for `R`, [e.g. Di Nardo et al., 2009].

For each simulation design we sample $S = 1000$ datasets and measure the accuracy of the estimates using the Frobenius distance d_F and the Amari

		$r=3$				$r=4$			
		$W_n = I_{d_g}$		$W_n = \widehat{\Sigma}_n^{-1}$		$W_n = I_{d_g}$		$W_n = \widehat{\Sigma}_n^{-1}$	
ε	d	d_F	d_A	d_F	d_A	d_F	d_A	d_F	d_A
1.	2	0.12	0.19	0.09	0.13	0.06	0.08	0.04	0.06
	3	0.12	0.25	0.08	0.18	0.07	0.15	0.05	0.10
2.	2	0.03	0.03	0.02	0.03	0.05	0.07	0.04	0.05
	3	0.03	0.06	0.02	0.05	0.06	0.12	0.04	0.08
3.	2	0.12	0.18	0.09	0.12	0.03	0.04	0.03	0.03
	3	0.11	0.23	0.13	0.33	0.03	0.07	0.03	0.06
4.	2	0.08	0.12	0.06	0.09	0.02	0.02	0.02	0.02
	3	0.07	0.17	0.09	0.21	0.02	0.04	0.02	0.04
5.	2	0.14	0.21	0.09	0.12	0.04	0.06	0.04	0.05
	3	0.14	0.27	0.12	0.25	0.05	0.09	0.04	0.08
6.	2	0.15	0.23	0.09	0.22	0.07	0.10	0.06	0.08
	3	0.14	0.28	0.12	0.24	0.07	0.15	0.06	0.12
7.	2	0.02	0.03	0.02	0.03	0.05	0.08	0.04	0.06
	3	0.02	0.05	0.02	0.04	0.06	0.13	0.05	0.10
8.	2	0.06	0.08	0.05	0.07	0.04	0.06	0.04	0.06
	3	0.06	0.12	0.05	0.11	0.04	0.10	0.04	0.10
9.	2	0.07	0.10	0.06	0.09	0.07	0.10	0.06	0.08
	3	0.06	0.12	0.06	0.12	0.07	0.14	0.06	0.13

TABLE 1. The table reports the average Frobenius norm (d_F) and Amari (d_A) errors obtain under diagonal cumulant restrictions for $r = 3, 4$ and weighting matrices $W_n = I_{d_g}$ and $W_n = \widehat{\Sigma}_n^{-1}$. The first column indicates the distribution from which the errors are sampled, see Figure 8.1, and the second column indicates the dimension of Y.

error d_A [e.g. Bach and Jordan, 2002a, Chen and Bickel, 2006]:

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

and

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

where a_{ij} is the i, j element of $A_0 \widehat{A}_{W_n}^{-1}$. We report the averages of these errors over the S datasets.

The results are shown in Table 8.1. A first key finding is that the estimates obtained using the optimal weighting matrix are nearly always more accurate when compared to taking $W_n = I_g$. This holds for $r = 3$ and $r = 4$, and for both the Frobenious and Amari measures of error. The only exception is found for the separated bi-modal densities 3. and 4. in the case where $r = 3$

and $d = 3$. This exception can be understood by noting third moments may not be very informative for symmetric bi-modal densities and the efficient weighting matrix cannot be estimated very accurately in these cases. In all other setting efficient weighting improves the outcomes, implying that conventional ICA algorithms can be modified in a simple way to increase efficiency. We note that in some setting the efficiency gains can be as large as 30-40%.

The differences between the orders $r = 3$ and $r = 4$ can be largely understood by reflecting on the underlying densities. For instance, for the skewed density 2. using the $r = 3$ order cumulant tensor restrictions gives more accurate estimates, whereas for Student's t density 1. the $r = 4$ order cumulant tensor is more accurate. Further, while the absolute accuracy as measured by the Amari error increases when we increase the dimension d, it is reassuring to see that for the averaged Frobinious norm measure, there are little differences when increasing d.

8.2. **Reflectionally invariant tensors.** Next, we evaluate the performance of the estimator (19) based on reflectionally invariant cumulant restrictions for order $r = 4$. For this setting we sample the errors from different transelliptical distributions such that the anti-symmetric monotone transformations $f(\varepsilon) = (\varepsilon_1^{1/3}, \ldots, \varepsilon_d^{1/3})$ are either: A. Student's t with $\nu = 15$, B. normal inverse Gaussian, C. Laplace and D. hyperbolic.

Under this design we have that (i) the components of ε are not independent, (ii) every even cumulant tensor is reflectionally invariant (cf. Proposition 5.7) and (iii) the genericity conditions in (9) hold. As such this design allows us to investigate the consequences of incorrectly assuming independence. Specifically, we compare with the estimator that relies on the (invalid) diagonal tensor $T \in S^4(\mathbb{R}^d)$, i,e. the JADE algorithm. For instance, for $d = 2$ the diagonal tensor sets incorrectly to zero $T_{1122} = 0$.

The results for both the baseline $W_n = I_{d_g}$ and efficient $W_n = \widehat{\Sigma}_n^{-1}$ cumulant estimators are shown in Table 8.2. The key finding is that invalid restrictions, e.g. incorrect independence assumptions, lead to substantial efficiency losses. Even when only one entry is incorrectly set to zero both the frobinious and Amari errors noticeably increase. This holds for all specifications and estimators considered.

Second, we find that in absolute terms the errors for the reflectionally invariant tensor are similar when compared to those found in Table 8.1. Third, the previous conclusions regarding the weighting matrix continue to hold; the efficient weighting matrix is nearly always preferable.

9. Discussion

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

		$r=4$, reflectional				$r=4$, diagonal			
		$W_n = I_{d_g}$		$W_n = \widehat{\Sigma}_n^{-1}$		$W_n = I_{d_g}$		$W_n = \widehat{\Sigma}_n^{-1}$	
ε	d	d_F	d_A	d_F	d_A	d_F	d_A	d_F	d_A
A.	2	0.03	0.05	0.03	0.04	0.05	0.08	0.04	0.08
	3	0.04	0.09	0.03	0.07	0.06	0.11	0.07	0.10
B.	2	0.05	0.07	0.04	0.07	0.07	0.09	0.07	0.10
	3	0.07	0.15	0.06	0.12	0.08	0.18	0.07	0.17
C.	2	0.06	0.08	0.05	0.07	0.08	0.10	0.08	0.10
	3	0.08	0.16	0.06	0.13	0.10	0.19	0.09	0.18
D.	2	0.06	0.08	0.05	0.08	0.08	0.11	0.08	0.11
	3	0.07	0.15	0.06	0.14	0.09	0.18	0.09	0.17

TABLE 2. The table reports the average Frobenius norm (d_F) and Amari (d_A) errors obtain under reflectionally invariant and (incorrect) diagonal cumulant restrictions for order $r = 4$ and weighting matrices $W_n = I_{d_g}$ and $W_n = \widehat{\Sigma}_n^{-1}$. The first column indicates the distributions from which the errors $f(\varepsilon)$, with anti-symmetric $f_i(\varepsilon) = \varepsilon_i^{1/3}$, are sampled: A. multivariate t-distribution 15 degrees of freedom, B. multivariate Normal Inverse Gaussian, C. multivariate Laplace and D. multivariate Hyperbolic distribution. The second column indicates the dimension of Y.

framework to study weak conditions under which A is identified up to some finite and structured set.

We propose some non-standard approaches 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 important zero patterns. These results can be used under strictly weaker conditions than independence and they assure that A is identified up to the sign permutations group acting on its rowspace. We note that with additional constraints, e.g. sign restrictions, these results can be further strengthened to exact identifiability.

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

ACKNOWLEDGEMENTS

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

References

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

Francis R. Bach and Michael I. Jordan. Tree-dependent component analysis. UAI'02, page 36–44, 2002b.

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.

David R Brillinger. The calculation of cumulants via conditioning. *Annals of the Institute of Statistical Mathematics*, 21(1):215–218, 1969.

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. Multidimensional independent component analysis. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 4, pages 1941–1944 vol.4, 1998.

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, December 1993.

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.

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.

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

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

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.

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

Economic Research, 2021.

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].

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.

Ronald Aylmer Fisher. Moments and product moments of sampling distributions. *Proceedings of the London Mathematical Society*, 2(1):199–238, 1930.

Gabriel Frahm, Markus Junker, and Alexander Szimayer. Elliptical copulas: applicability and limitations. *Statistics & Probability Letters*, 63(3):275–286, 2003.

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.

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.

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

Aapo Hyvärinen and Patrik Hoyer. Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Computing*, 12:1705–20, 2000.

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

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.

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.

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.

Han Liu, Fang Han, and Cun-hui Zhang. Transelliptical graphical models. *Advances in neural information processing systems*, 25, 2012.

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.

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.

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.

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.

Hannu Oja, Davy Paindaveine, and Sara Taskinen. Affine-invariant rank tests for multivariate independence in independent component models. *Electronic Journal of Statistics*, 10(2):2372 – 2419, 2016.

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.

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

V. P. Skitivic. 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.

Terry P. Speed. Cumulants and partition lattices ii: Generalised k-statistics. *Journal of the Australian Mathematical Society*, 40(1):34–53, 1986.

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

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.

Appendix A. Moments and Cumulants

For the reader's convenience, in this appendix we collect some further standard results on moments and cumulants and their sample estimates that are used in this paper.

A.1. **Combinatorial relationship between moments and cumulants.** Let $\mathbf{\Pi}_r$ be the poset of all set partitions of $\{1,\dots,r\}$ ordered by refinement. For $\pi \in \mathbf{\Pi}_r$ 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 $\boldsymbol{i} = (i_1,\dots,i_r)$ then $\boldsymbol{i}_B$ is a subvector of $\boldsymbol{i}$ 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].

$$[\mu_r(Y)]_{i_1,\dots,i_r} = \sum_{\pi\in\Pi_r} \prod_{B\in\pi} [\kappa_{|B|}(Y)]_{\boldsymbol{i}_B} \ , \tag{22}$$

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

$$[\mu_r(Y)]_{i_1,i_2,i_3} = \kappa_{i_1i_2i_3} + \kappa_{i_1i_2}\kappa_{i_3} + \kappa_{i_1i_3}\kappa_{i_2} + \kappa_{i_2i_3}\kappa_{i_1} + \kappa_{i_1}\kappa_{i_2}\kappa_{i_3} \ ,$$

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

$$[\kappa_r(Y)]_{i_1,\dots,i_r} = \sum_{\pi\in\Pi_r} (-1)^{|\pi|-1}(|\pi|-1)! \prod_{B\in\pi} [\mu_{|B|}(Y)]_{\boldsymbol{i}_B} \ . \tag{23}$$

For example,

$$[\kappa_r(Y)]_{i_1,i_2,i_3} \ = \ \mu_{i_1i_2i_3} - \mu_{i_1}\mu_{i_2i_3} - \mu_{i_2}\mu_{i_1i_3} - \mu_{i_3}\mu_{i_1i_2} + 2\mu_{i_1}\mu_{i_2}\mu_{i_3} \ ,$$

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

The coefficients $(-1)^{|\pi|-1}(|\pi|-1)!$ in (23) have an important combinatorial interpretation, which we now briefly explain. If $\boldsymbol{P}$ is a finite partially ordered set (poset) with ordering $\leq$ we define the zeta function on $\boldsymbol{P}\times\boldsymbol{P}$ as $\zeta(x,y) = 1$ if $x \leq y$ and $\zeta(x,y) = 0$ otherwise. The Möbius function is then defined by setting $\mathfrak{m}(x,y) = 0$ if $x \not\leq y$ and

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

Fixing a total ordering on $\boldsymbol{P}$, 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 $\boldsymbol{P}$ 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)

$$\mathfrak{m}(\rho, \pi) \;=\; (-1)^{|\rho|-|\pi|} \prod_{B \in \pi} (|\rho_B| - 1)!, \tag{24}$$

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 $\pi \in \mathbf{\Pi}_r$

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

To explain how $\mathfrak{m}(\pi, \mathbf{1})$ appears in (23), we recall the Möbius inversion formula, which becomes clear given the matrix formulation using Z and $M = Z^{-1}$.

Lemma A.1 (Möbius inversion theorem)**.** *Let $\boldsymbol{P}$ be a poset. For two functions c, d on $\boldsymbol{P}$, we have $d(x) = \sum_{y \leq x} c(y)$ for all $x \in \boldsymbol{P}$ if and only if $c(x) = \sum_{y \leq x} \mathfrak{m}(x, y) d(y)$.*

For example, this result gives the simple formula (22) that defines moments in terms of cumulants.

A.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 A.2 (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*

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

where for $\boldsymbol{i} = (i_1, \ldots, i_r)$

$$\left[\operatorname{cum}((\kappa_{|B|}(X|H))_{B \in \pi})\right]_{\boldsymbol{i}} = \operatorname{cum}\left((\operatorname{cum}(X_{\boldsymbol{i}_B}|H))_{B \in \pi}\right).$$

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

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

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

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

Proposition A.2 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.

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

$$[\hat{\boldsymbol{\mu}}_r]_{i_1\dots i_r} = \frac{1}{n}\sum_{s=1}^{n} Y_{s,i_1}Y_{s,i_2}\dots Y_{s,i_r}\ . \tag{25}$$

Using our multilinear notation we can more compactly write

$$\hat{\boldsymbol{\mu}}_r = \frac{1}{n}\boldsymbol{Y}'\bullet I_r\ \in\ S^r(\mathbb{R}^d). \tag{26}$$

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

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 rth 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)])

$$[\mathsf{k}_r]_{i_1,\dots,i_r} = \frac{1}{n}\sum_{t_1=1}^{n}\cdots\sum_{t_r=1}^{n}\Phi_{t_1,\dots,t_r}Y_{t_1,i_1}\cdots Y_{t_r,i_r} \tag{27}$$

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

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

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

$$\mathsf{k}_r\ =\ \frac{1}{n}\boldsymbol{Y}'\bullet\Phi\ \in\ S^r(\mathbb{R}^d). \tag{28}$$

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

Proposition A.3. *The* k*-statistic in* (27) *forms a U-statistic. In particular, it is an unbiased and it has the minimal variance among all unbiased estimators.*

Besides being unbiased and efficient, an additional benefit of working with 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].

A.4. **k-statistics and sample cumulants.** For later considerations we need to understand better the relation between k_r and the natural plug-in estimator $\hat{\boldsymbol{\kappa}}_r$, which is obtained by first estimating the raw moments and then plugging them into (23). The relevant sample moments that allow to compute $\hat{\boldsymbol{\kappa}}_r$ from (23) 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

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

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

$$c(\pi) = \sum_{\rho\leq\pi} \mathfrak{m}(\rho,\pi)(-1)^{|\rho|-1}\frac{1}{\binom{n-1}{|\rho|-1}} = n\sum_{\rho\leq\pi}\mathfrak{m}(\rho,\pi)\mathfrak{m}(\rho,\mathbf{1})\frac{1}{(n)_{|\rho|}}, \tag{29}$$

where $\mathfrak{m}$ is the Möbius function on $\mathbf{\Pi}_r$ given in (24) and $(n)_k = n(n-1)\cdots(n-k+1)$ is the corresponding falling factor.

Lemma A.4. *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

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

Proof. For any $t_1,\ldots,t_r$ 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

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

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 A.1.

The second claim follows from the fact that

$$\mathsf{k}_r \overset{(28)}{=} \frac{1}{n}\boldsymbol{Y}'\bullet\Phi = \frac{1}{n}\sum_{\pi\in\mathbf{\Pi}_r} c(\pi)\bigotimes_{B\in\pi}(\boldsymbol{Y}'\bullet I_B) \overset{(26)}{=} \sum_{\pi\in\mathbf{\Pi}_r} n^{|\pi|-1}c(\pi)\bigotimes_{B\in\pi}\hat{\boldsymbol{\mu}}_B,$$

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

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

Lemma A.5. *For every $\pi \in \mathbf{\Pi}_r$ we have*

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

Proof. As we noted in the proof of Lemma A.4, the Möbius inversion formula in Lemma A.1 gives that

$$\sum_{\rho\leq\pi} c(\rho) = (-1)^{|\pi|-1}\frac{1}{\binom{n-1}{n-|\pi|}}. \tag{30}$$

Let $\mathbf{0} \in \mathbf{\Pi}_r$ be the minimal partition into r singleton blocks. By (30), applied to $\pi = \mathbf{0}$,

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

where $(n)_r = n\cdots(n-r+1)$ is the corresponding falling factorial. In particular, $n^{r-1}c(\mathbf{0}) = \mathfrak{m}(\mathbf{0},\mathbf{1}) + O(n^{-1})$. 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 $n^{|\rho|-1}c(\rho) = \mathfrak{m}(\rho,\mathbf{1}) + O(n^{-1})$ so

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

This assures that

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

Using (30) in the same way as above, we get that $n^{|\pi|-1}c(\pi) = \mathfrak{m}(\pi,\mathbf{1}) + O(n^{-1})$ and now the result follows by recursion. □

A.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 $\mathrm{vec}(T) \in \mathbb{R}^{d^r}$ 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 with no need to discuss inverses of quadratic forms.

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

$$\mathrm{vec}(S) \;=\; A^{\otimes r}\cdot\mathrm{vec}(T), \tag{31}$$

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

A.6. **Asymptotic distribution of sample statistics.** To derive the asymptotic distribution of the minimum distance estimators in Section 7 we require the asymptotic distribution of the sample moments or the k-statistics.

Sample moments

The sample moments of Y are defined as in (25). When using moments the distance measure $\hat{m}_n(A)$ (see (16)) depends on the tensors $\hat{\boldsymbol{\mu}}_2$ and $\hat{\boldsymbol{\mu}}_r$.

As formalized in the lemma below, we have that under suitable moment assumptions that

$$\hat{\boldsymbol{\mu}}_p \xrightarrow{p} \mu_p(Y) \qquad \forall\ p \leq r\ , \tag{32}$$

and

$$\sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\mu}}_2 - \mu_2(Y), \hat{\boldsymbol{\mu}}_r - \mu_r(Y)) \xrightarrow{d} N(0, V)\ , \tag{33}$$

where V is the asymptotic variance matrix with entries

$$V_{\boldsymbol{i},\boldsymbol{j}} = \mathrm{cov}(Y_{i_1} \cdots Y_{i_k}, Y_{i_1} \cdots Y_{i_l}) \qquad k, l \in \{2, r\}\ .$$

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 (33) we can use (31) to derive the limiting distribution of $\hat{m}_n(A)$ for moments. We have

$$\begin{aligned}\sqrt{n}\mathrm{vec}(\hat{m}_n(A) - m(A)) &= [A^{\otimes 2}, A^{\otimes r}] \cdot \sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\mu}}_2 - \mu_2(Y), \hat{\boldsymbol{\mu}}_r - \mu_r(Y)) \\ &\xrightarrow{d} N(0, A^{2,r} V A^{2,r'})\end{aligned} \tag{34}$$

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

$$\Sigma_\mu^{2,r} = A^{2,r} V A^{2,r'}\ . \tag{35}$$

k-statistics

Next, we provide analog steps for the k-statistics. First, let $\boldsymbol{\mu}_{\leq r}$ be the vector containing all moments of a random vector Y of order up to r (it has dimension $\binom{d+r}{r}$). Formula (23) 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 $F = \nabla_{\boldsymbol{\mu}'_{\leq r}} \mathrm{vec}(\kappa_r(Y))$, 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 $1 \leq i_1 \leq \cdots \leq i_r \leq d$ and $1 \leq j_1 \leq \cdots \leq j_r \leq d$ we have that

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

and so, this submatrix contains the identity matrix.

Under suitable moment conditions we have

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

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

As in Appendix A.4, denote $\hat{\boldsymbol{\kappa}}_r$ to be the image of $\hat{\boldsymbol{\mu}}_{\leq r}$ under the map (23). It then follows from the delta method that

$$\sqrt{n}\,\mathrm{vec}(\hat{\boldsymbol{\kappa}}_r - \kappa_r(Y)) \xrightarrow{d} N(0, FHF')\ . \tag{36}$$

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 (36). We will show that k-statistics k_r have the same asymptotic distribution.

Lemma A.6. *If* $\mathbb{E}\|Y_s\|^{2r} < \infty$ *we have that*

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

Proof. By (36) and Slutsky lemma, it is enough to show that $\sqrt{n}\,(\mathsf{k}_r - \hat{\boldsymbol{\kappa}}_r) \xrightarrow{p} 0$. By Lemma A.4,

$$[\mathsf{k}_r - \hat{\boldsymbol{\kappa}}_r]_{i_1\cdots i_r} \ = \ \sum_{\pi\in\mathbf{\Pi}_r} (n^{|\pi|-1}c(\pi) - \mathfrak{m}(\pi,\mathbf{1})) \prod_{B\in\pi} \hat{\boldsymbol{\mu}}_{\boldsymbol{i}_B},$$

where the coefficients $c(\pi)$ are defined in (29). By Lemma A.5, $n^{|\pi|-1}c(\pi) - \mathfrak{m}(\pi,\mathbf{1}) \ = \ O(n^{-1})$ for all $\pi \in \mathbf{\Pi}_r$ and so in particular

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

Under the stated moment assumption $\hat{\boldsymbol{\mu}}_{\boldsymbol{i}_B} = O_p(1)$ and so $[\mathsf{k}_r - \hat{\boldsymbol{\kappa}}_r]_{i_1\cdots i_r} = o_P(1)$, which completes the proof. □

By Lemma A.6, every linear transformation of $\sqrt{n}\mathrm{vec}(\mathsf{k}_r - \kappa_r(Y))$ will be also Gaussian. We will be in particular interested in transformations $A^{\otimes r}\mathrm{vec}(\mathsf{k}_r - \kappa_r(Y))$ as motivated by the multilinear action of A on $S^r(\mathbb{R}^d)$ (cf. (31)). We have

$$\sqrt{n}A^{\otimes r}\mathrm{vec}(\mathsf{k}_r - \kappa_r(Y)) \quad \xrightarrow{d} \quad N(0, A^{\otimes r}FH(A^{\otimes r}F)') \ .$$

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 $\mathrm{vec}(\kappa_2(Y), \kappa_r(Y)) \in \mathbb{R}^{d^2+d^r}$. By exactly the same arguments as above we get

$$\sqrt{n}\,\mathrm{vec}(\mathsf{k}_2 - \kappa_2(Y), \mathsf{k}_r - \kappa_r(Y)) \quad \xrightarrow{d} \quad N(0, F^{2,r}H(F^{2,r})') \ . \tag{37}$$

Recall from (16) 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 k-statistics in Section 7.

$$\mathrm{vec}(\hat{m}_n(A) - m(A)) \ = \ [A^{\otimes 2}, A^{\otimes r}] \cdot \mathrm{vec}(\mathsf{k}_2 - \kappa_2(Y), \mathsf{k}_r - \kappa_r(Y)). \tag{38}$$

We will write $A^{2,r} = [A^{\otimes 2}, A^{\otimes r}]$ and, using (37), we immediately conclude

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

Let this asymptotic covariance matrix be denoted by

$$\Sigma^{2,r}_{\mathsf{k}} = A^{2,r}F^{2,r}H(A^{2,r}F^{2,r})'. \tag{39}$$

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

Lemma A.7. *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* $\mathsf{k}_p - \kappa_p(Y) \xrightarrow{p} 0$ *for all* $p \leq r$.

(2) *if* $\mathbb{E}\|Y_s\|^{2r} < \infty$, *then*

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

where $h = \mu$ *or* $h = \mathsf{k}$ *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* (35) *and the cumulant based variance* $\Sigma_{\mathsf{k}}^{2,r}$ *in* (39).

Appendix B. Specific cumulant calculations

B.1. **Scale mixture of normals.** Since the distribution of spherical distributions is invariant under orthogonal transformations, in particular all its cumulants must be reflectionally invariant. In this section we make some specific calculations specializing to *scale mixtures of normals*, which forms a big subfamily of elliptical distributions. For these distributions we have the following stochastic representation:

$$X = \mu + \frac{1}{\sqrt{\tau}} \cdot Z \ , \tag{40}$$

where τ is a positive random variable with no atom at the origin, $Z \sim N(0, \Sigma)$ and $\tau \perp\!\!\!\perp Z$. The normal distribution corresponds to $\tau \equiv 1$. If $\tau \sim \chi^2_\nu/\nu$ for $\nu > 2$ then X has a multivariate t-distribution. If $\nu = 1$ we get the multivariate Cauchy, and if $\tau \sim \mathrm{Exp}(1)$ the multivariate Laplace distribution. The following lemma defines the cumulant tensors for X.

Lemma B.1. *If* X *has the scale mixture of normals distribution* (40) *with* $\mu = 0$ *then* $\kappa_2(X) = \mathbb{E}(\frac{1}{\tau})\Sigma$, $\kappa_r(X) = \mathbf{0}$ *if* r *is odd, and, if* $r = 2l$,

$$(\kappa_r(X))_{i_1 \cdots i_r} = \kappa_l(\tfrac{1}{\tau}) \sum_{j_1 k_1 / \ldots / j_l k_l} \Sigma_{j_1 k_1} \cdots \Sigma_{j_l k_l} \ ,$$

where the sum runs over all two-block partitions of the set $\{1, \ldots, r\}$ *with* $\{j_1, k_1, \ldots, j_l, k_l\} = \{1, \ldots, r\}$.

Proof. Use the stochastic representation (40) of X in terms of a Gaussian Z. It follows that the conditional distribution of X given τ is Gaussian with mean μ and covariance $\frac{1}{\tau}\Sigma$. Recall also that all cumulants of a mean zero Gaussian vector are zero apart from the second order cumulants (covariances). Denote by $\mathbf{\Pi}_r^{(2)}$ the set of partitions in $\mathbf{\Pi}_r$ with all blocks precisely of size 2. If r is odd then $\mathbf{\Pi}_r^{(2)} = \emptyset$. By the law of total cumulants in Proposition A.2, conditioning X on τ we get

$$[\kappa_r(X)]_{i_1 \cdots i_r} = \sum_{j_1 k_1 / \cdots / j_l k_l \in \mathbf{\Pi}_r^{(2)}} \mathrm{cum}(\mathrm{cov}(X_{j_1}, X_{k_1}|\tau), \ldots, \mathrm{cov}(X_{j_l}, X_{k_l}|\tau)).$$

We have $\operatorname{cov}(X_j, X_k|\tau) = \frac{1}{\tau}\Sigma_{jk}$ and so, using also multilinearity of cumulants, the above formula further simplifies

$$[\kappa_r(X)]_{i_1\cdots i_r} = \sum_{j_1k_1/\cdots/j_lk_l \in \mathbf{\Pi}_r^{(2)}} \Sigma_{j_1k_1}\cdots\Sigma_{j_lk_l}\kappa_r(1/\tau)$$

giving the final formula. □

A special case is obtained for $\Sigma = I_d$, where the formula in Lemma B.1 further simplifies. For a given $\boldsymbol{i} = (i_1 \cdots i_r)$, let n_j be the number of times the index j appeared in $\boldsymbol{i}$. Then

$$(\kappa_r(X))_{i_1\cdots i_r} = \kappa_r(\tfrac{1}{\tau}) \begin{cases} 0 & \text{if some } n_j \text{ is odd,} \\ (n_1-1)!!\cdots(n_d-1)!! & \text{otherwise} \end{cases}$$

with a convention that $(-1)!! = 1$. Note that so defined tensor $\kappa_r(X)$ is isotropic in the sense that $Q \bullet \kappa_r(X) = \kappa_r(X)$ for all $Q \in O(d)$. This tensor is *not* generic in the sense of Theorem 5.9.

B.2. **Cumulants Gaussian mixture.**

Lemma B.2. *Denote $\Delta = \Sigma_2 - \Sigma_1$. If X is a mixture of zero-mean Gaussians with parameters γ, Σ_1, Σ_2 then all the odd-order cumulants are zero. If $r = 2l \geq 4$ we have*

$$(\kappa_r(X))_{i_1\ldots i_r} = \kappa_l(H) \cdot \sum_{j_1k_1/\cdots/j_lk_l} \Delta_{j_1k_1}\cdots\Delta_{j_lk_l}.$$

Proof. Conditionally on H the variable X has a mean zero Gaussian distribution and the only non-zero conditional cumulants are the conditional covariances. We have

$$\operatorname{cov}(X_j, X_k|H) \;=\; (1-H)(\Sigma_1)_{jk} + H(\Sigma_2)_{jk} \;=\; (\Sigma_1)_{jk} + H\Delta_{jk}.$$

As in the proof of Lemma B.1 we use the law of total cumulance to conclude that

$$[\kappa_r(X)]_{i_1\cdots i_r} = \sum_{j_1k_1/\cdots/j_lk_l \in \mathbf{\Pi}_r^{(2)}} \operatorname{cum}(\operatorname{cov}(X_{j_1}, X_{k_1}|H), \ldots, \operatorname{cov}(X_{j_l}, X_{k_l}|H)).$$

If $l \geq 2$ ($r \geq 2$) we can use the invariance of cumulants under translations to simplify this to the claimed form. □

We are now ready to prove Theorem 2.4.

Proof of Theorem 2.4. If the components of X are independent then $\kappa = \kappa_4(X)$ is a diagonal tensor by Proposition 3.2. If $\gamma \in \{0, 1\}$ this clearly holds. But in this case the distribution of X is Gaussian. Suppose $\gamma \in (0, 1)$ so that $\kappa_2(H) > 0$. By Lemma B.2, the condition $\kappa_{ijjj} = \kappa_{iijj} = \kappa_{iiij} = 0$ for all $i < j$ already implies that

$$\Delta_{ii}\Delta_{ij} = \Delta_{jj}\Delta_{ij} = \Delta_{ii}\Delta_{jj} + 2\Delta_{ij}^2 = 0. \tag{41}$$

The only way this holds for all $i \neq j$ is that either $\Delta = \mathbf{0}$ or $\Delta_{ii} \neq 0$ for some i but otherwise Δ is zero. Indeed, if $\Delta_{ij} = 0$ then the first two constraints in (41) imply that $\Delta_{ii} = \Delta_{jj} = 0$ but then the third constraint in (41) cannot hold. We conclude that Δ is diagonal. The third constraint in (41) shows then that at most one diagonal entry can be non-zero.

If Δ is zero then the conditional distribution of ε given H does not depend on H and so it is Gaussian. Finally, consider the case when $\delta = \Delta_{11} > 0$ with the remaining entries of Δ zero. Since this is a non-trivial mixture of Gaussian distributions, at least one component of X must be non-Gaussian. Note however that the distribution of $(X_2, \ldots, X_d)$ does not depend on H and so it is Gaussian. It follows that exactly one component of X is non-Gaussian. Moreover, by the independence of the components of X and by the law of total covariance, for each $i \neq j$

$$0 = \operatorname{cov}(X_i, X_i) = \mathbb{E}(\operatorname{cov}(X_i, X_j|H)) = (\Sigma_1)_{ij} + \gamma\Delta_{ij} = (\Sigma_1)_{ij}.$$

This proves that both Σ_1 and Σ_2 must be diagonal.

□

If we do not assume independence things are more interesting. In the context of non-independent component analysis it is natural to assume that ε is a mixture of two zero-mean Gaussian distributions with diagonal matrices. If Y is a mixture of $\mathcal{N}(0, \Sigma_1)$ and $\mathcal{N}(0, \Sigma_2)$, this is equivalent to assuming that both Σ_1 and Σ_2 are simultaneously diagonalizable by the same orthogonal matrix.

Lemma B.3. *Suppose that ε is a mixture of zero-mean Gaussian distributions with diagonal Δ. Then all odd-order cumulants are zero and $(\kappa_r(\varepsilon))_{\boldsymbol{i}} \neq 0$ for even r only if all indices in $\boldsymbol{i} = (i_1, \ldots, i_r)$ appear even number of times. In particular, $\kappa_r(\varepsilon)$ must be reflectionally invariant.*

Proof. Follows immediately from Lemma B.2. □

B.3. **Mean independence in the binary case.** In this section we prove Proposition 5.14. The condition $Q \bullet T \in \mathcal{V}$ translates into two equations $(Q \bullet T)_{12\cdots 2} = (Q \bullet T)_{1\cdots 12} = 0$. In other words,

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

and

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

where in both cases the sum goes over all $(r-1)$-tuples $\boldsymbol{j}$. Note that, since T is symmetric, the entry $T_{\boldsymbol{i}}$ depends only on how many times 1 appears in $\boldsymbol{i}$. Write $t_k = T_{\boldsymbol{i}}$ if $\boldsymbol{i}$ has k ones. 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

$$\sum_{k=0}^{r-1}\binom{r-1}{k}z^{k}t_{k+1}-\sum_{k=0}^{r-1}\binom{r-1}{k}z^{k+1}t_{k} \;=\; 0 \tag{42}$$

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

$$\sum_{k=0}^{r-1}\binom{r-1}{k}(-1)^{k}z^{k+1}t_{r-k}+\sum_{k=0}^{r-1}\binom{r-1}{k}(-1)^{k}z^{k}t_{r-k-1} \;=\; 0. \tag{43}$$

Using the fact that $t_1=t_{r-1}=0$, (42) 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 (43) 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

$$a_k=\binom{r-1}{k+1}t_{k+2}-\binom{r-1}{k}t_{k}\qquad\text{for } k=0,\ldots,r-2$$

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 $Q \bullet T \in \mathcal{V}$.

Appendix C. Omitted proofs from Section 7

C.1. **Proof of Proposition 7.3.** 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 C.1. *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 7.3. First, note that $\widehat{A}_{W_n}$ minimizes $\hat{L}_{W_n}(A)$ and we take $L_W(A)$ as $L_0(\theta)$ in Theorem C.1. 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 $Q \in \mathrm{SP}(d)$, 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 $Q \in \mathrm{SP}(d)$. 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 C.2. *Suppose that* $\{Y_s\}_{s=1}^n$ *is i.i.d,* $W_n \xrightarrow{p} W$, $\mathbb{E}\|Y_s\|^r < \infty$, *and* $\mathcal{A} \subset \mathrm{GL}(d)$ *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 $\|\hat{\boldsymbol{\mu}}_p - \mu_p(Y)\| \xrightarrow{p} 0$ and $\|\mathsf{k}_p - \kappa_r(Y)\| \xrightarrow{p} 0$ for any $p \le r$ by Lemma A.7 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 7 we have that $\|\hat{h}_p - h_p(Y)\| \xrightarrow{p} 0$ for $p \le r$. Hence,

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

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 (38), we get

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

As $g_{S,T}(A)$ is defined in (17) 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

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

The second term is 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

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

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

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

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

We may now apply Theorem C.1 to conclude that $\widehat{A}_{W_n} \xrightarrow{p} QA_0$ for some $Q \in \mathrm{SP}(d)$.

C.2. **Proof of Proposition 7.4.** 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 C.3. *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) $\theta_0 \in \mathrm{Int}(\Theta)$, (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,*

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

The loss $\hat{L}_n(\theta)$ in Theorem C.3 corresponds to our $\hat{L}_n(A)$. Our $\hat{g}_n(A)$ corresponds to their $\hat{g}_n(\theta)$. We have $\widehat{A}_{W_n} \xrightarrow{p} \tilde{A}_0 = QA_0$ for some $Q \in \mathrm{SP}(d)$ by Proposition 7.3, and the conditions on the weighting matrix are satisfied by (ii). Condition (a) of Theorem C.3 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 A.7, $\sqrt{n}\,\mathrm{vec}(\hat{m}_n(\tilde{A}_0) - m(\tilde{A}_0))$ weakly converges to $N(0, \Sigma_h^{2,r})$, where $h = \mu$ or $h = \mathsf{k}$ pending whether moments or k-statistics are used to compute $\hat{m}_n(A)$. The variance matrices are defined in (35) or (39). 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

$$\Sigma = D_{\mathcal{I}}^{2,r} \Sigma_h^{2,r} D_{\mathcal{I}}^{2,r'} \tag{44}$$

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 D 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 (17) 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 (11), and the projection $\pi_{\mathcal{V}}$. Thus, the derivative is given by mapping $V \in \mathbb{R}^{d \times d}$ to the vector

$$\mathrm{vec}\Big((V, A) \bullet S + (A, V) \bullet S, \ \pi_{\mathcal{V}}\big((V, A, \ldots, A) \bullet T + \cdots + (A, \ldots, A, V) \bullet T\big)\Big).$$

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 $(E_{ij} \otimes A + A \otimes E_{ij})\mathrm{vec}(S)$ over the vector

$$\big((E_{ij} \otimes A \otimes \cdots \otimes A) + \cdots + (A \otimes \cdots \otimes A \otimes E_{ij})\big) \cdot \mathrm{vec}(T), \tag{45}$$

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 $\widehat{G}(A)$ if $S = \hat{\boldsymbol{\mu}}_2$, $T = \hat{\boldsymbol{\mu}}_r$ (or $S = \kappa_2(Y)$, $T = \kappa_r(Y)$ and $S = \mathsf{k}_2$, $T = \mathsf{k}_r$). The columns of $\widehat{G}(A) - G(A)$ are like explained in (45) with $S = \hat{\boldsymbol{\mu}}_2 - \mu_2(Y)$ and $T = \hat{\boldsymbol{\mu}}_r - \mu_r(Y)$ (or $S = \mathsf{k}_2 - \kappa_2(Y)$ and $T = \mathsf{k}_r - \kappa_r(Y)$). Since $\|S\| \xrightarrow{p} 0$ and $\|T\| \xrightarrow{p} 0$ by Lemma A.7 part 1, and because A is fixed, the norm of each column converges to zero. In consequence, for each A, $\|\widehat{G}(A) - G(A)\| \xrightarrow{p} 0$. Since $\mathcal{A}$ is compact and $\widehat{G}(A) - G(A)$ is smooth, we conclude

$$\sup_{A \in \mathcal{A}} \|\widehat{G}(A) - G(A)\| \xrightarrow{p} 0. \tag{46}$$

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 C.4 below.

Lemma C.4. *If $\mathcal{V}$ assures identifiability up to a sign permutation matrix, then the matrix $G(QA_0)$ has full column rank for each $Q \in \mathrm{SP}(d)$.*

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 (11), 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 6.5 and the fact that I_d is an isolated point of $\mathcal{G}_T$. We conclude that V must be zero. □

Having verified all conditions of C.3 we can apply the theorem to prove the first display in Proposition 7.4. The second display follows as a special case when taking $W_n = \widehat{\Sigma}_n^{-1}$, noting that $\widehat{\Sigma}_n^{-1} \to \Sigma^{-1}$, and replacing W by Σ^{-1} in the first display.

C.3. **Proof of Proposition 7.5.** Let $\tilde{A}_0 = QA_0$. Noting that $\hat{g}_n(\widehat{A}_{\widehat{\Sigma}_n^{-1}})$ minimizes $\|\cdot\|^2_{W_n}$ when taking $W_n = \widehat{\Sigma}_n^{-1}$, we get that $\hat{g}_n(\widehat{A}_{\widehat{\Sigma}_n^{-1}}) = 0$. Using Taylor's theorem we get that

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

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

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

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

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

where

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

By the discussion preceding (44), 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 $\widehat{\Sigma}_n^{-1/2}\sqrt{n}\hat{g}_n(\tilde{A}_0) \xrightarrow{d} Z \sim N(0, I_{d_g})$ only by something that converges to zero in probability, as $\widehat{\Sigma}_n \xrightarrow{p} \Sigma$. By Slutsky's lemma we have $\widehat{\Sigma}^{-1/2}\sqrt{n}\hat{g}_n(\tilde{A}_0) \xrightarrow{d} Z \sim N(0, I_{d_g})$, and from Proposition 7.3, equation (46) and $\widehat{\Sigma}_n \xrightarrow{p} \Sigma$ and the continuous mapping theorem, we get

$$\widehat{N} \xrightarrow{p} N = I_{d_g} - \Sigma^{-1/2}G(\tilde{A}_0)[G(\tilde{A}_0)'\Sigma^{-1}G(\tilde{A}_0)]^{-1}G(\tilde{A}_0)'\Sigma^{-1/2} \ . \tag{47}$$

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

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

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

C.4. **Proof of Proposition 7.6.** From the proof of Proposition 7.5 we have

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

where are N is the projection matrix defined in (47). 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

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

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 $\operatorname{rank}(N-B) = \operatorname{Tr}(N-B) = d_g - d_{g_1}$. To complete the proof note that

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

Appendix D. Computing the asymptotic variance

In this section we give computational details for estimating the asymptotic variance matrices Σ and S as defined in Proposition 7.4. Starting with Σ, 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

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

and for cumulants

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

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 $\operatorname{vec}(\hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\mu}}_r)$ and $\hat{\boldsymbol{\mu}}_{\leq r}$, respectively, $A^{2,r} = [A^{\otimes 2}, A^{\otimes r}]$ and $F^{2,r}$ is the Jacobian matrix of the transformation from $\boldsymbol{\mu}_{\leq r}$ to cumulants (κ_2, κ_r).

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 $\hat{\boldsymbol{\mu}}_r$. Further, $A^{2,r} = [A^{\otimes 2}, A^{\otimes r}]$ can replaced by its estimate $\widehat{A}_{W_n}^{2,r} = [\widehat{A}_{W_n}^{\otimes 2}, \widehat{A}_{W_n}^{\otimes r}]$. Combining we obtain the estimates

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

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. Fortunately it is easy to see that Σ_h can be also estimated using a simple bootstrap. Let $\hat{\boldsymbol{\varepsilon}}_n = \widehat{A}_{W_n}\mathbf{Y}_n$ denote the $n\times 1$ vector of residuals. We can resample these residuals (with replacement) to get $\hat{\boldsymbol{\varepsilon}}_n^*$ and construct bootstrap draws of $\hat{g}_n(\widehat{A}_{W_n})$, say g_n^*. Repeating this B times allows to compute the bootstrap variance estimate

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

The $1/n$ comes from the definition $\Sigma = \lim_{n\to\infty}\operatorname{var}(\sqrt{n}\hat{g}_n(QA_0))$. 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 $\hat{\boldsymbol{\varepsilon}}_n^*$.

While the bootstrap is conceptually attractive, it is worth nothing that, at least in principle, the covariance between two k-statistics $\mathsf{k}_{i_1\cdots i_r}$ and $\mathsf{k}_{j_1\cdots j_r}$ can be computed exactly for any given sample size using the general formula

for cumulants of 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 (28) to obtain the covariance in any given case by noting that

$$\mathbb{E}\mathrm{vec}(\mathsf{k}_r)\mathrm{vec}(\mathsf{k}_r)' \;=\; \frac{1}{n^2}\mathbb{E}\left[(\boldsymbol{Y}')^{\otimes r}\mathrm{vec}(\Phi)\mathrm{vec}(\Phi)'\boldsymbol{Y}^{\otimes r}\right].$$

Note however that $\mathrm{vec}(\Phi)$ 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 $\widehat{A}_{W_n}$ and the map (45) provides the estimate for G. Combining this an estimate for Σ as defined above allows to estimate S.

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