SUPPLEMENTARY MATERIAL: NON-INDEPENDENT COMPONENTS ANALYSIS

BY GEERT MESTERS[1,a], PIOTR ZWIERNIK[2,b]

[1]Department of Economics and Business, Universitat Pompeu Fabra, [a]geert.mesters@upf.edu

[2]Department of Statistical Sciences, University or Toronto, [b]piotr.zwiernik@utoronto.ca

We provide the following additional results.

S0. Omitted proofs – main text.

In this section we collect the omitted proofs from the main text.

S0.1. *Omitted proof from Section 5 .*

PROOF OF PROPOSITION 5.18. 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 $\boldsymbol{j} = (j_1,\ldots,j_{r-1}) \in \{1,2\}^{r-1}$. 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, $k = 0,\ldots,r$. With this notation the two equations above simplify to

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

and

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

If one of the entries of Q is zero then Q is a permutation matrix. So assume that Q has no zeros. Assume also without loss of generality that Q is a rotation matrix, that is, $Q_{11} = Q_{22}$

MSC2020 subject classifications: Primary 15A69, 62H99; secondary .
Keywords and phrases: independent component analysis, identifiability, cumulants, tensors.

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

$$\text{(S1)} \qquad \sum_{k=0}^{r-1} \binom{r-1}{k} z^k t_{k+1} - \sum_{k=0}^{r-1} \binom{r-1}{k} z^{k+1} t_k \;=\; 0$$

and

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

It is convenient to rewrite the latter as

$$\text{(S2)} \qquad \sum_{k=0}^{r-1} \binom{r-1}{k} (-1)^k z^{k+1} t_{r-k} + \sum_{k=0}^{r-1} \binom{r-1}{k} (-1)^k z^k t_{r-k-1} \;=\; 0.$$

Using the fact that $t_1 = t_{r-1} = 0$, (S1) 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 (S2) 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 recursively 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 $Q \bullet T \in \mathcal{V}$. □

S0.2. *Omitted proofs from Section 6.*

PROOF OF LEMMA 6.1. We have $L_W(A) = 0$ if and only if $g(A) = 0$, which is equivalent $A \bullet h_2(Y) = I_d$ and $A \bullet h_r(Y) \in \mathcal{V}$. Since (1) holds, we also have $A_0 \bullet h_2(Y) = I_2$ and $A_0 \bullet h_r(Y) \in \mathcal{V}$. It follows that $A_0^{-1}A \in \mathrm{O}(d)$, or in other words, $A = QA_0$ for some $Q \in \mathrm{O}(d)$. Further,

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

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

PROOF OF PROPOSITION 6.2. The proof follows from verifying the conditions for consistency of a general extremum estimator. Specifically, we will verify the conditions of Theorem 2.1 in Newey and McFadden (1994). We restate the theorem for completeness.

THEOREM S1. *Suppose that $\hat{\theta}$ minimizes $\hat{L}_n(\theta)$ over $\theta \in \Theta$. Assume that there exists a function $L_0(\theta)$ such that (a) $L_0(\theta)$ is uniquely minimized at θ_0, (b) $L_0(\theta)$ is continuous, (c) Θ is compact and (d) $\sup_{\theta\in\Theta} |\hat{L}_n(\theta) - L_0(\theta)| \xrightarrow{p} 0$, then $\hat{\theta} \xrightarrow{p} \theta_0$.*

Next, we verify assumptions (a)-(d) under assumptions (i)-(iv) stated in Proposition 6.2. First, note that $\widehat{A}_{W_n}$ minimizes $\hat{L}_{W_n}(A)$ and we take $L_W(A)$ as $L_0(\theta)$ in Theorem S1. Second, in our case the minimizer of $L_W(A)$ is not unique but will correspond to any of the finite points QA_0 for some $Q \in SP(d)$. It follows that our consistency result will only be up to permutation and sign changes of the true A_0 (e.g. Chen and Bickel, 2006). Formally, for (a): suppose that A is such that $A \neq QA_0$ for any $Q \in \mathrm{SP}(d)$, 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 S2. *Suppose that $\{Y_s\}_{s=1}^n$ is i.i.d, $W_n \xrightarrow{p} W$, $\mathbb{E}\|Y_s\|^r < \infty$, and $\mathcal{A} \subset \mathrm{GL}(d)$ 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 $p \leq r$ by Lemma S7 part 1. Note that the norm $\|\cdot\|$ on the tensor is defined in the usual way as the sum of the squares of all elements. Using the general notation of Section 6 we have that $\|\hat{h}_p - h_p(Y)\| \xrightarrow{p} 0$ for $p \leq r$. Hence,

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

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

Using (S28), we get

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

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 S1 to conclude that $\widehat{A}_{W_n} \xrightarrow{p} QA_0$ for some $Q\in\mathrm{SP}(d)$. □

S0.3. *Proof of Proposition 6.3.* The proof follows from verifying the conditions for asymptotic normality of a generalized moment or distance estimator. Specifically, we will verify the conditions of Theorem 3.2 in Newey and McFadden (1994). We restate the theorem for completeness.

THEOREM S3. *Suppose that $\hat\theta$ minimizes $\hat{L}_n(\theta)$ over $\theta\in\Theta$ with Θ compact, where $\hat{L}_n(\theta)$ is of the form $\hat{g}_n(\theta)'W_n\hat{g}_n(\theta)$ and $W_n\xrightarrow{p}W$ with W positive semi-definite, $\hat\theta\xrightarrow{p}\theta_0$ and (a) $\theta_0\in\mathrm{Int}(\Theta)$, (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 S3 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 6.2, and the conditions on the weighting matrix are satisfied by (ii). Condition (a) of Theorem S3 is satisfied by assumption (v). For (b) note that $\hat{g}_n(A)$ is a polynomial map in A and hence smooth. For (c), by Lemma S7, $\sqrt{n}\,\mathrm{vec}(\hat{m}_n(\tilde{A}_0)-m(\tilde{A}_0))$ 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 (S25) or (S29). 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

$$\Sigma = D_{\mathcal{I}}^{2,r}\Sigma_h^{2,r}D_{\mathcal{I}}^{2,r'} \tag{S3}$$

with $D^{2,r}_{\mathcal{I}}$ 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 S5 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 (S6), and the projection $\pi_{\mathcal{V}}$. Thus, the derivative is given by mapping $V \in \mathbb{R}^{d\times d}$ to the vector

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

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

$$\big((E_{ij} \otimes A \otimes \cdots \otimes A) + \cdots + (A \otimes \cdots \otimes A \otimes E_{ij})\big) \cdot \mathrm{vec}(T), \tag{S4}$$

and then selecting only the entries corresponding to the 2-tuples $i \le j$ 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 (S4) 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 S7 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

$$\sup_{A\in\mathcal{A}} \|\widehat{G}(A) - G(A)\| \xrightarrow{p} 0. \tag{S5}$$

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

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

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 (S6), we write this last condition as $K_{I_d,I_d}(A_0^{-1}Q'V) = 0$. Similarly, the $\mathcal{V}^{\perp}$-part implies that $K_{T,I_d}(A_0^{-1}Q'V) = 0$ with $T = \kappa_r(Y)$. This implies that $A_0^{-1}Q'V = 0$ by Lemma S5 and the fact that I_d is an isolated point of $\mathcal{G}_T$. We conclude that V must be zero. □

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

S1. Additional motivation. In this section we discuss some additional relations that aim to further highlight the usefulness of the identification results presented in the main text.

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

Now reconsider the multiple scaled elliptical components model

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

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

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

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

S1.2. *Invariance.* In Section 2 we motivated non-independent component models using specific examples (e.g. common variance model) as well as by relaxing independence (e.g. mean independence). In both cases the resulting model still implied sufficient zero restrictions on the higher order moments/cumulants of ε to ensure the identifiability of A (cf Corollaries 5.7 and 5.15). Here we briefly show that such zero restrictions can also arise from invariance properties of the distribution of ε.

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

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

Since D is arbitrary, $[\kappa_r(\varepsilon)]_{i_1 \cdots i_r}$ 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}. These zero patterns correspond exactly with the reflectionally invariant restrictions introduced in Section 5.2 and as such Corollary 5.15 also ensure the identifiability of A in $AY = \varepsilon$ when the distribution of ε is the same as the distribution of $D\varepsilon$.

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

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

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

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

S2. Local identification beyond signed-permutations. The results in Section 5 stipulate conditions on moment tensors $T = \mu_r(\varepsilon)$ or cumulant tensors $T = \kappa_r(\varepsilon)$ for which A can be recovered up to sign and permutation. This section gives minimal conditions on $\mathcal{V}$ that ensure that $\mathcal{G}_T$ is finite. We subsequently use this result to highlight the gap that exists between restrictions that lead to finite sets and restrictions that lead to signed permutation sets. This finding has the important implication that it is in general not sufficient to prove that the Jacobian of the moment or cumulant restrictions is full rank in order to establish that the identified set is equal to the set of signed permutations.

Let $\mathcal{V} \subset S^r(\mathbb{R}^d)$ be a set given as a set of zeros of a system of polynomials in the coordinated of $S^r(\mathbb{R}^d)$ (such set is called an algebraic variety). A subset $\mathcal{U} \subseteq \mathcal{V}$ is Zariski open in $\mathcal{V}$ if the complement $\mathcal{V} \setminus \mathcal{U}$ is also an algebraic variety. In particular, a Zariski open set is also open in the classical topology. For example, the set of diagonal tensors in $S^r(\mathbb{R}^d)$ with at most one zero on the diagonal forms a Zariski open subset of the set of diagonal tensors. Similarly, the set of reflectionally invariant tensors satisfying the genericity condition (14) is Zariski open in the set of reflectionally invariant tensors. Note that, in both cases, the constraints defining $\mathcal{V}$ and $\mathcal{V} \setminus \mathcal{U}$ were linear.

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

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

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

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

PROOF. The right implication is clear. For the left implication first note that $\mathcal{G}_{T^*}(\mathcal{U})$ is a Zariski open subset of the real algebraic variety $\mathcal{G}_{T^*}(\mathcal{V})$. Indeed, if $f_1(T) = \cdots = f_k(T) = 0$ are the polynomials, in T, describing $\mathcal{V}$ then the polynomials, in Q, 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, Little and OShea (2013). It is then enough to show that if Q° is isolated in $\mathcal{G}_{T^*}(\mathcal{U})$ then it must be isolated in $\mathcal{G}_{T^*}(\mathcal{V})$. Suppose that $Q^\circ \in \mathcal{G}_{T^*}(\mathcal{U})$ is not isolated in $\mathcal{G}_{T^*}(\mathcal{V})$. Then it must lie on an irreducible component of the variety $\mathcal{G}_{T^*}(\mathcal{V})$ of a positive dimension. By assumption, for this Q°, $g_1(Q^\circ \bullet T^*) \neq 0, \ldots, g_l(Q^\circ \bullet T^*) \neq 0$. Thus, in any sufficiently small neighbourhood of Q° there will be a point that lies in $\mathcal{G}_{T^*}(\mathcal{V})$ and $g_1, \dots, g_l$ evaluate to something non-zero. In other words, in any sufficiently small neighbourhood of Q° there is a point in $\mathcal{G}_{T^*}(\mathcal{U})$ proving that Q° cannot be isolated in $\mathcal{G}_{T^*}(\mathcal{U})$, which leads to contradiction. □

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

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

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

$$K_{T,A}(V) \;=\; (V, A, \dots, A) \bullet T + \cdots + (A, \dots, A, V) \bullet T. \tag{S6}$$

Moreover, if A is invertible, then

$$K_{T,A}(V) \;=\; K_{A\bullet T, I_d}(VA^{-1}). \tag{S7}$$

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

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

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 (S6) for $r = 2$.

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

$$K_{I_d,Q}(V) = 0 \text{ and } \pi_{\mathcal{V}}(K_{T,Q}(V)) = 0 \qquad \text{implies } V = 0. \tag{S8}$$

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

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

Taking the limit $t \to 0$, 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 S6. In the examples of Section 5, for $T \in \mathcal{U} \subseteq \mathcal{V}$, we always had $\mathcal{G}_T(\mathcal{U}) = \mathcal{G}_T(\mathcal{V}) = \mathrm{SP}(d)$. The proof of Proposition S2 suggests that, at least in principle $\mathcal{G}_T(\mathcal{U})$ could be finite but $\mathcal{G}_T(\mathcal{V})$ could have components of positive dimension. In the proof of the next result, we crucially rely on the fact that we compute $\mathcal{G}_T(\mathcal{U})$ rather than $\mathcal{G}_T(\mathcal{V})$.

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

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

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

$$B^{(j)} \;=\; [T_{klj\cdots j}]_{k,l<j} \;\in\; S^2(\mathbb{R}^{j-1}) \tag{S9}$$

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

$$\det\left(T_{j\cdots j}I_{j-1} - (r-1)B^{(j)}\right) \neq 0 \quad \text{for all } j = 2,\ldots,d. \tag{S10}$$

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

PROOF. By Proposition S2 it is enough to show that each point of $\mathcal{G}_T(\mathcal{U}^\circ)$ is isolated. By Lemma S5, equivalently for every $Q \in \mathcal{G}_T(\mathcal{U}^\circ)$, if $K_{I_d,Q}(V) = 0$ and $\pi_{\mathcal{V}}(K_{T,Q}(V)) = 0$ then $V = 0$. By (S7), $K_{I_d,Q}(V) \;=\; K_{I_d,I_d}(VQ')$. 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 (S7), 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 (S9) 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 (S10). 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 $j \geq 2$. We will use the fact that $U_{jl} = 0$ if $l \geq j$. 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 (S10). Using a recursive argument, we conclude that $U = 0$. □

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

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

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 S9. The set $\mathcal{G}_T(\mathcal{U}^\circ)$ is finite but, as illustrated by Example S8, it typically contains matrices that do not have an easy interpretation. In particular, if $d = 2$ then $\mathcal{V}^\circ$ 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 S10. *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.*

S3. Moments and Cumulants — some useful properties. We collect some results on moments and cumulants and their sample estimates that are used below for some of the proofs.

S3.1. *Combinatorial relationship between moments and cumulants.* Let $\boldsymbol{\Pi}_r$ be the poset of all set partitions of $\{1,\ldots,r\}$ ordered by refinement. For $\pi \in \boldsymbol{\Pi}_r$ 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 $\boldsymbol{i} = (i_1,\ldots,i_r)$ then $\boldsymbol{i}_B$ is a subvector of $\boldsymbol{i}$ with indices corresponding to the block $B \subseteq \{1,\ldots,r\}$. For any multiset $\{i_1,\ldots,i_r\}$ of the indices $\{1,\ldots,d\}$ we can relate the moments $\mu_r(Y)$ to the cumulants (e.g. Speed, 1983).

$$[\mu_r(Y)]_{i_1,\ldots,i_r} = \sum_{\pi\in\Pi_r}\prod_{B\in\pi}[\kappa_{|B|}(Y)]_{\boldsymbol{i}_B}\,, \tag{S11}$$

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

$$[\kappa_r(Y)]_{i_1,\dots,i_r} = \sum_{\pi\in\Pi_r} (-1)^{|\pi|-1}(|\pi|-1)! \prod_{B\in\pi} [\mu_{|B|}(Y)]_{\boldsymbol{i}_B}\ . \tag{S12}$$

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 (S12) 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 $\boldsymbol{\Pi}_r$ the Möbius function satisfies for any $\rho\leq\pi$ (ρ is a refinement of π)

$$\mathfrak{m}(\rho,\pi) \ = \ (-1)^{|\rho|-|\pi|} \prod_{B\in\pi} (|\rho_B|-1)!, \tag{S13}$$

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

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

To explain how $\mathfrak{m}(\pi,\boldsymbol{1})$ appears in (S12), we recall the Möbius inversion formula, which becomes clear given the matrix formulation using Z and $M=Z^{-1}$.

LEMMA S1 (Möbius inversion theorem). *Let $\boldsymbol{P}$ 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 (S11) that defines moments in terms of cumulants.

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

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

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

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

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

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

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

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

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

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

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

$$[\hat{\boldsymbol{\mu}}_r]_{i_1\ldots i_r} = \frac{1}{n}\sum_{s=1}^n Y_{s,i_1}Y_{s,i_2}\ldots Y_{s,i_r}\ . \tag{S14}$$

Using our multilinear notation we can more compactly write

$$\hat{\boldsymbol{\mu}}_r = \frac{1}{n}\boldsymbol{Y}'\bullet I_r \,\in\, S^r(\mathbb{R}^d). \tag{S15}$$

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 $[\kappa_r(Y)]_{i_1\ldots i_r}$ are given by (see (McCullagh, 2018, (4.5)-(4.7)))

$$[\mathsf{k}_r]_{i_1,\ldots,i_r} = \frac{1}{n}\sum_{t_1=1}^n\cdots\sum_{t_r=1}^n \Phi_{t_1,\ldots,t_r}Y_{t_1,i_1}\cdots Y_{t_r,i_r} \tag{S16}$$

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 $(t_1,\ldots,t_n)$. Let $\boldsymbol{Y}\in\mathbb{R}^{n\times d}$ be the data matrix. More compactly, we have

$$\mathsf{k}_r \;=\; \frac{1}{n}\boldsymbol{Y}'\bullet\Phi \,\in\, S^r(\mathbb{R}^d). \tag{S17}$$

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

PROPOSITION S3. *The* k*-statistic in* (S16) *forms a U-statistic. In particular, it is unbiased and it has the minimal variance among all unbiased estimators.*

Besides being unbiased and efficient, an additional benefit of working with 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, Guarino and Senato (2009).

S3.4. *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 (S12). The relevant sample moments that allow to compute $\hat{\boldsymbol{\kappa}}_r$ from (S12) are summarized in $\hat{\boldsymbol{\mu}}_p$ for $p \le r$.

If $B \subseteq [n]$ then write I_B for the identity tensor in $S^{|B|}(\mathbb{R}^n)$. For any partition $\pi \in \boldsymbol{\Pi}_r$ 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 $\pi \in \boldsymbol{\Pi}_r$, define coefficients

$$c(\pi) = \sum_{\rho\le\pi} \mathfrak{m}(\rho,\pi)(-1)^{|\rho|-1}\frac{1}{\binom{n-1}{|\rho|-1}} = n\sum_{\rho\le\pi}\mathfrak{m}(\rho,\pi)\mathfrak{m}(\rho,\mathbf{1})\frac{1}{(n)_{|\rho|}}, \tag{S18}$$

where $\mathfrak{m}$ is the Möbius function on $\boldsymbol{\Pi}_r$ given in (S13) and $(n)_k = n(n-1)\cdots(n-k+1)$ is the corresponding falling factor.

LEMMA S4. *We have*

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

which gives an alternative formula for k-statistics

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

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

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

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 $\boldsymbol{\Pi}_r$ as given in Lemma S1.

The second claim follows from the fact that

$$\mathsf{k}_r \overset{(S17)}{=} \frac{1}{n}\boldsymbol{Y}'\bullet\Phi = \frac{1}{n}\sum_{\pi\in\boldsymbol{\Pi}_r} c(\pi)\bigotimes_{B\in\pi}(\boldsymbol{Y}'\bullet I_B) \overset{(S15)}{=} \sum_{\pi\in\boldsymbol{\Pi}_r} n^{|\pi|-1}c(\pi)\bigotimes_{B\in\pi}\hat{\boldsymbol{\mu}}_B,$$

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 S5. *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 S4, the Möbius inversion formula in Lemma S1 gives that

$$\sum_{\rho \le \pi} c(\rho) = (-1)^{|\pi|-1} \frac{1}{\binom{n-1}{n-|\pi|}}. \tag{S19}$$

Let $\mathbf{0} \in \mathbf{\Pi}_r$ be the minimal partition into r singleton blocks. By (S19), 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

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

Using (S19) 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. □

S3.5. *Vectorizations of tensors.* The dimension of the space of symmetric tensors $S^r(\mathbb{R}^d)$ is $\binom{d+r-1}{r}$. Like for symmetric matrices, it is often convenient to view $T \in S^r(\mathbb{R}^d)$ as a general tensor in $\mathbb{R}^{d \times \cdots \times d}$. In this case $\mathrm{vec}(T) \in \mathbb{R}^{d^r}$ 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 without referring to abstract linear algebra.

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

$$\mathrm{vec}(S) \;=\; A^{\otimes r} \cdot \mathrm{vec}(T), \tag{S20}$$

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

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

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

$$m_{S,T}(A) \;=\; (A \bullet S - I_d,\; A \bullet T)\;. \tag{S21}$$

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 $\hat{\boldsymbol{\mu}}_r$, or the rth order k-statistic, denoted by k_r, which are computed from a given sample $\{Y_s\}_{s=1}^n$ as discussed above. It is worth pointing out that these results generalize existing results (e.g. Jammalamadaka, Taufer and Terdik, 2021) for the asymptotic analysis of cumulant estimates to higher order tensors.

Sample moments

The sample moments of Y are defined as in (S14). When using moments the distance measure $\hat{m}_n(A)$ (see (S21)) 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

$$\hat{\boldsymbol{\mu}}_p \xrightarrow{p} \mu_p(Y) \qquad \forall\, p \le r , \tag{S22}$$

and

$$\sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\mu}}_2 - \mu_2(Y), \hat{\boldsymbol{\mu}}_r - \mu_r(Y)) \xrightarrow{d} N(0, V) , \tag{S23}$$

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 (S23) we can use (S20) to derive the limiting distribution of $\hat{m}_n(A)$ for moments. We have

$$\begin{aligned} \sqrt{n}\mathrm{vec}(\hat{m}_n(A) - m(A)) &= [A^{\otimes 2}, A^{\otimes r}] \cdot \sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\mu}}_2 - \mu_2(Y), \hat{\boldsymbol{\mu}}_r - \mu_r(Y)) \\ &\xrightarrow{d} N(0, A^{2,r} V A^{2,r'}) \end{aligned} \tag{S24}$$

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

$$\Sigma_\mu^{2,r} = A^{2,r} V A^{2,r'} . \tag{S25}$$

k-statistics

Next, we provide analog steps for the k-statistics. First, let $\boldsymbol{\mu}_{\le r}$ be the vector containing all moments of a random vector Y of order up to r (it has dimension $\binom{d+r}{r}$). Formula (S12) gives an explicit function for $\kappa_r(Y)$ in terms of $\boldsymbol{\mu}_{\le r}$. For the vectorized tensor $\kappa_r(Y)$ we define the Jacobian $F = \nabla_{\boldsymbol{\mu}'_{\le r}}\mathrm{vec}(\kappa_r(Y))$, 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 $1 \le i_1 \le \cdots \le i_r \le d$ and $1 \le j_1 \le \cdots \le j_r \le d$ we have that

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

and so, this submatrix contains the identity matrix.

Under suitable moment conditions we have

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

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

As in Appendix S3.4, denote $\hat{\boldsymbol{\kappa}}_r$ to be the image of $\hat{\boldsymbol{\mu}}_{\leq r}$ under the map (S12). It then follows from the delta method that

$$\sqrt{n}\,\mathrm{vec}(\hat{\boldsymbol{\kappa}}_r - \kappa_r(Y)) \;\xrightarrow{d}\; N(0, FHF')\,. \tag{S26}$$

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 (S26). We will show that k-statistics k_r have the same asymptotic distribution.

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

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

Proof. By (S26) and Slutsky lemma, it is enough to show that $\sqrt{n}\,(\mathsf{k}_r - \hat{\boldsymbol{\kappa}}_r) \xrightarrow{p} 0$. By Lemma S4,

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

where the coefficients $c(\pi)$ are defined in (S18). By Lemma S5, $n^{|\pi|-1}c(\pi) - \mathfrak{m}(\pi,\mathbf{1}) = O(n^{-1})$ for all $\pi \in \boldsymbol{\Pi}_r$ 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 S6, 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. (S20)). We have

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

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

$$\sqrt{n}\,\mathrm{vec}(\mathsf{k}_2 - \kappa_2(Y), \mathsf{k}_r - \kappa_r(Y)) \;\xrightarrow{d}\; N(0, F^{2,r}H(F^{2,r})')\,. \tag{S27}$$

Recall from (S21) 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 6.

$$\mathrm{vec}(\hat{m}_n(A) - m(A)) \;=\; [A^{\otimes 2}, A^{\otimes r}] \cdot \mathrm{vec}(\mathsf{k}_2 - \kappa_2(Y), \mathsf{k}_r - \kappa_r(Y)). \tag{S28}$$

We will write $A^{2,r} = [A^{\otimes 2}, A^{\otimes r}]$ and, using (S27), we immediately conclude

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

Let this asymptotic covariance matrix be denoted by

$$\Sigma_{\mathsf{k}}^{2,r} = A^{2,r}F^{2,r}H(A^{2,r}F^{2,r})'. \tag{S29}$$

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

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

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

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

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 (S25) and the cumulant based variance $\Sigma_{\mathsf{k}}^{2,r}$ in (S29).

S4. Additional inference tools. In this section we complement the inference Section 6 with some additional tools that can be used to select the appropriate moment/cumulant zero restrictions in a data driven way.

S4.1. *Testing over-identifying restrictions.* While zero restrictions on higher order moments or cumulants can be motivated from several angles (cf. the discussion in Section 2), it is useful to test ex-post whether the restrictions indeed appear to hold in a given application. In the setting where d_g is strictly greater 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 S1. *If the conditions of Proposition 6.3 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 S1 as a general test for model misspecification.

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

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

PROPOSITION S2. *If the conditions of Proposition 6.3 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.

S5. Computing the asymptotic variance. In this section we give computational details for estimating the asymptotic variance matrices Σ and S as defined in Proposition 6.3. Starting with Σ (see equation (21)) 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

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

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 $\mathrm{vec}(\hat{\boldsymbol{\mu}}_2,\hat{\boldsymbol{\mu}}_r)$ and $\hat{\boldsymbol{\mu}}_{\le r}$, respectively, $A^{2,r}=[A^{\otimes 2},A^{\otimes r}]$ and $F^{2,r}$ is the Jacobian matrix of the transformation from $\boldsymbol{\mu}_{\le r}$ to cumulants (κ_2,κ_r), see Section S3.6 for explicit definitions.

The moment matrices V and H and the Jacobian matrix $F^{2,r}$ can be estimated by replacing the population moments of $\mu_r(Y)$ by the sample moments $\hat{\boldsymbol{\mu}}_r$. 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

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

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

Therefore, for cumulant restriction we recommend estimating Σ_h using a simple bootstrap. Let $\hat{\varepsilon}_n=\widehat{A}_{W_n}\mathbf{Y}_n$ denote the $n\times 1$ vector of residuals. We can resample these residuals (with replacement) to get $\hat{\varepsilon}_n^*$ 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

$$\widehat{\Sigma}_n/n = \frac{1}{B}\sum_{b=1}^{B}(g_n^{*,b}-\bar{g}_n^*)(g_n^{*,b'}-\bar{g}_n^*)' \qquad \text{with} \quad \bar{g}_n^* = \frac{1}{B}\sum_{b=1}^{B}g_n^{*,b}\,.$$

The $1/n$ comes from the definition $\Sigma=\lim_{n\to\infty}\mathrm{var}(\sqrt{n}\hat{g}_n(QA_0))$. Using the bootstrap has the benefit that no additional analytical calculations are needed and evaluating $g_n^{*,b}$ only requires computing the sample statistics $\mu_p(Y)$ or k_p, for $p=2,r$, for each bootstrap draw $\hat{\varepsilon}_n^*$. The validity of the bootstrap follows as we have a random sample $\{Y_i\}$, $\widehat{A}_{W_n}$ is $\sqrt{n}$-consistent for A and asymptotically normal (cf Propositions 6.2 and 6.3) and $\hat{g}_n(A)$ is a polynomial map in A and hence smooth.

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 (S17) 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 (S4) provides the estimate for G. Combining this an estimate for Σ as defined above allows to estimate S.

S6. Additional numerical results.

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

S6.1. *Alternative performance measures.* We start by providing the same results as in the main text but now measuring the accuracy of the different procedures in terms of the Frobenius distance d_F (e.g. Chen and Bickel, 2006) which is often referred to as the minimum distance index (e.g. Ilmonen et al., 2010) and can be defined as

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

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

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

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

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

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

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

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

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

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

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

TABLE S1
MINIMUM DISTANCE: COMMON VARIANCE MODEL

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

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

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

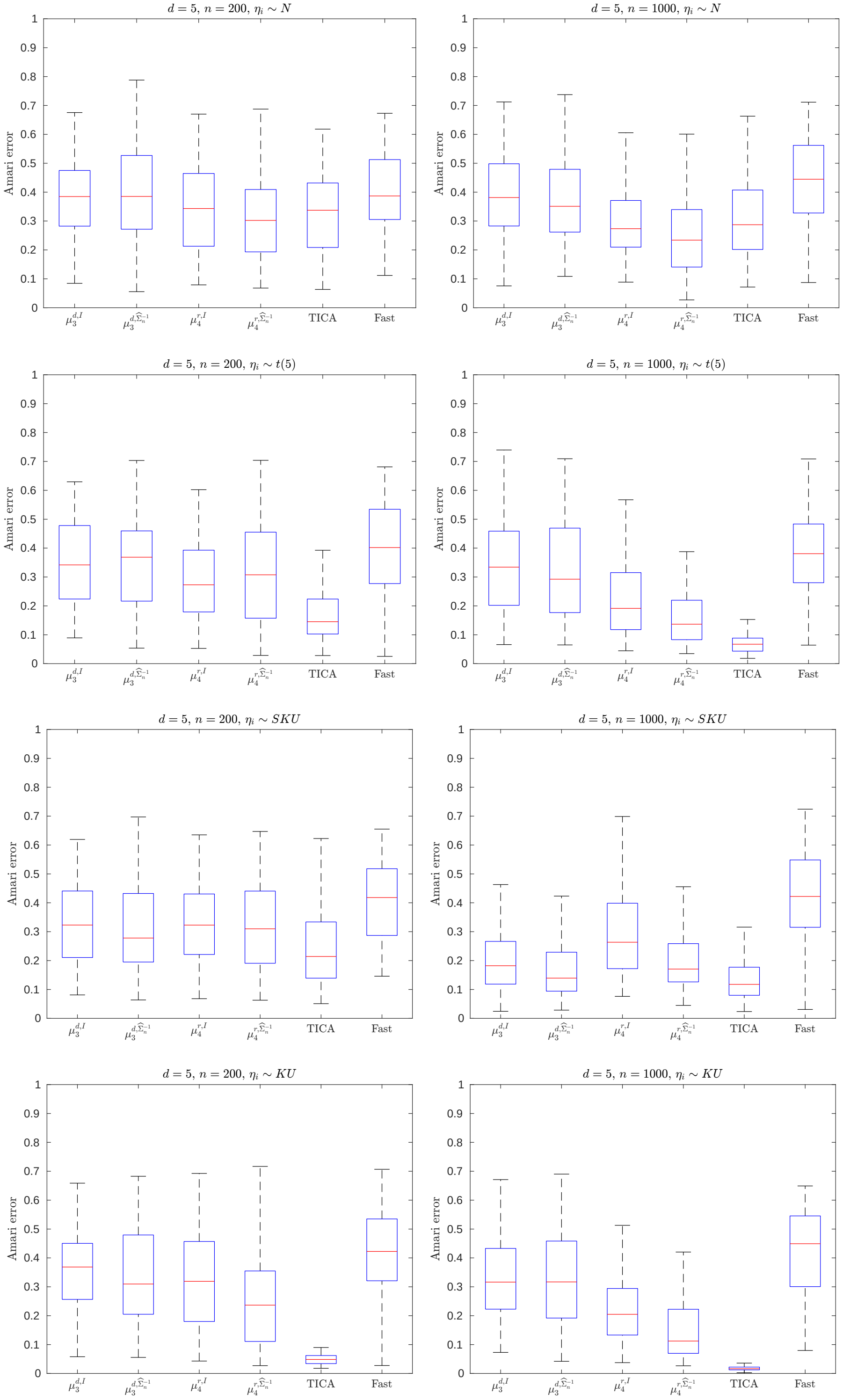
TABLE S2
MINIMUM DISTANCE: SCALED ELLIPTICAL

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

	Independent Components Analysis									
Method	$\mathcal{N}$	$t(5)$	SKU	KU	BM	SBM	SKB	TRI	CL	ACL
Fast	0.14	0.13	0.14	0.14	0.13	0.14	0.14	0.14	0.13	0.14
JADE	0.14	0.14	0.14	0.14	0.14	0.14	0.15	0.14	0.14	0.14
Kernel	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
ProDen	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.15
Efficient	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
NPML	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14

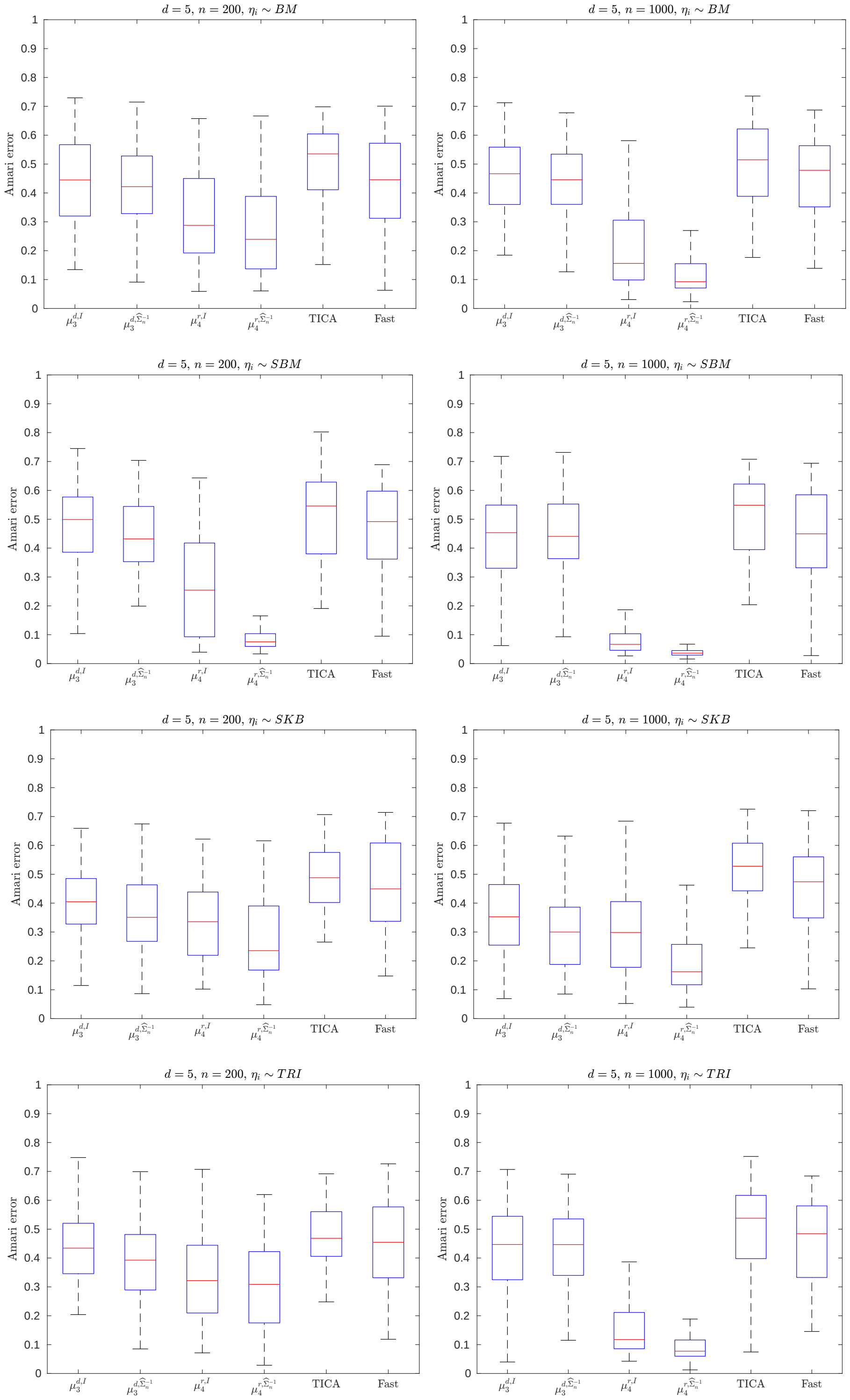
Notes: The table reports the average Minimum Distance errors (across $S = 1000$ simulations) for data sampled from the multiple scaled elliptical model (4) with $d = 2$ and $n = 200$. The columns correspond to the different errors considered for the components of η, see Table 1. The top panel reports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

Fig S1: Common Variance Experiments



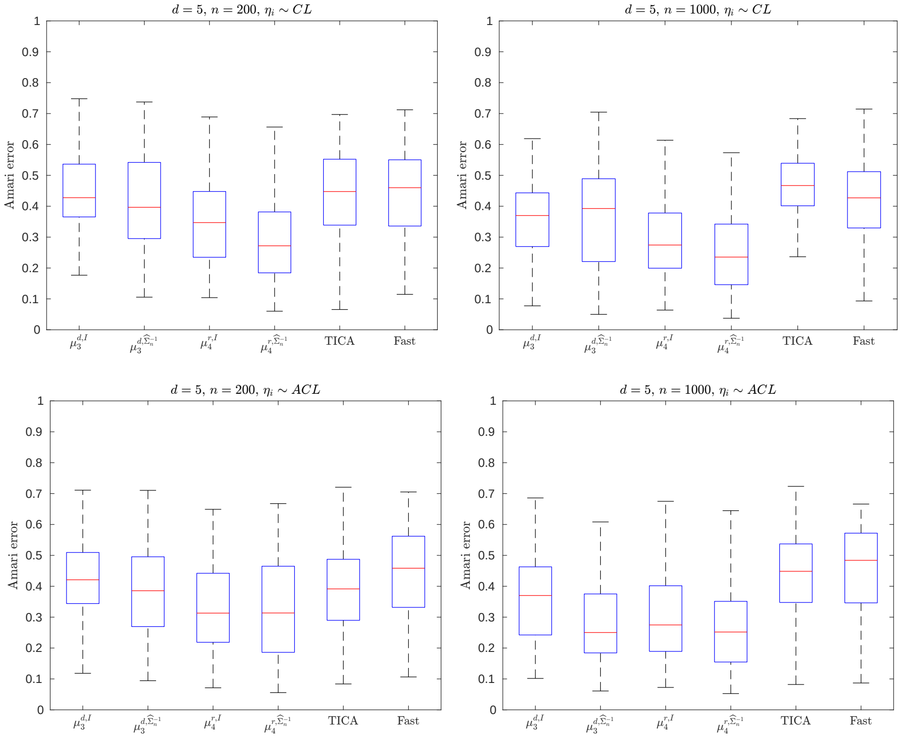
Notes: The figure shows the boxplots for the Amari errors (across $S = 100$ simulations) for data sampled from the common variance model (3). The different settings for the simulations designs are described in the titles and the x-labels indicate the different estimation methods used.

Fig S2: Common Variance Experiments



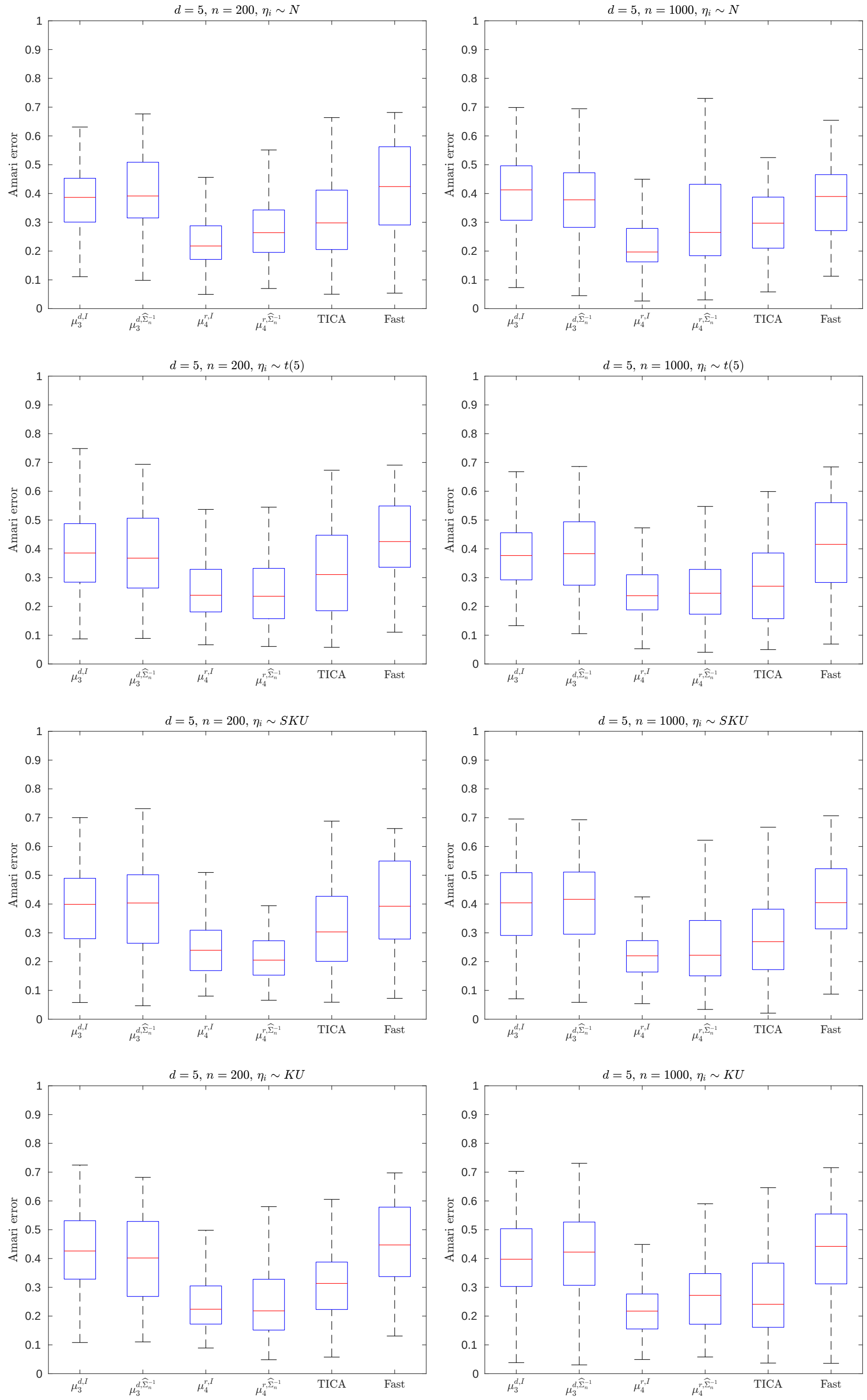
Notes: The figure shows the boxplots for the Amari errors (across $S = 100$ simulations) for data sampled from the common variance model (3). The different settings for the simulations designs are described in the titles and the x-labels indicate the different estimation methods used.

Fig S3: Common Variance Experiments



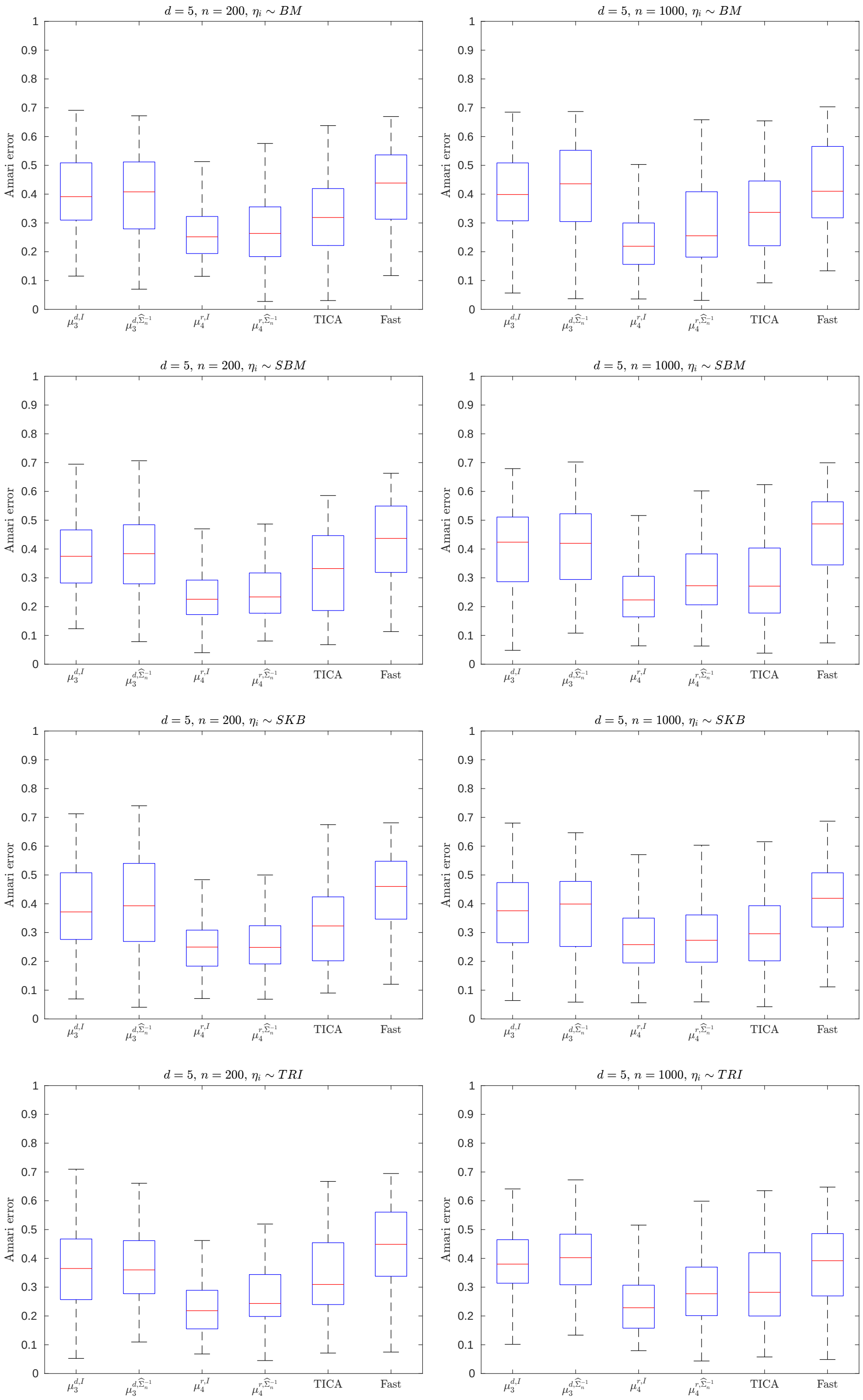
Notes: The figure shows the boxplots for the Amari errors (across $S = 100$ simulations) for data sampled from the common variance model (3). The different settings for the simulations designs are described in the titles and the x-labels indicate the different estimation methods used.

Fig S4: Scaled Elliptical Experiments



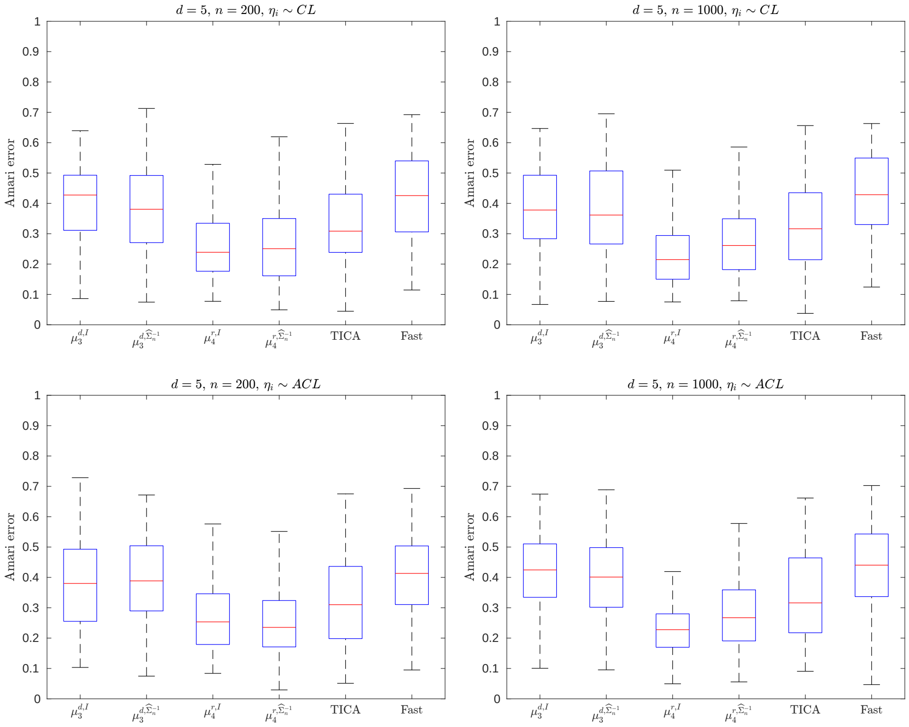
Notes: The figure shows the boxplots for the Amari errors (across $S = 100$ simulations) for data sampled from the multiple scaled elliptical components model (4). The different settings for the simulations designs are described in the titles and the x-labels indicate the different estimation methods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

Fig S5: Scaled Elliptical Experiments



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

Fig S6: Scaled Elliptical Experiments



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

S7. Omitted proofs.

S7.1. *Proof of Proposition S1.* 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 (S3), 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 6.2, equation (S5) and $\widehat{\Sigma}_n \xrightarrow{p} \Sigma$ and the continuous mapping theorem, we get

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

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).

S7.2. *Proof of Proposition S2.* From the proof of Proposition S1 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 (S30). Let $\hat{g}_{1,n}$, G_1, N_1 be the equivalent quantities to $\hat{g}_n$, G, N just computed for the smaller set of identifying restrictions. Using similar arguments we get

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

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 $N_1 \Xi P = N_1 P_1 = 0$, 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 $\mathrm{rank}(N - B) = \mathrm{Tr}(N - B) = d_g - d_{g_1}$. To complete the proof note that

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

REFERENCES

BACH, F. R. and JORDAN, M. I. (2003). Beyond Independent Components: Trees and Clusters. *Journal of Machine Learning Research* **4** 1205–1233.

BRILLINGER, D. R. (1969). The calculation of cumulants via conditioning. *Annals of the Institute of Statistical Mathematics* **21** 215–218.

CARDOSO, J.-F. and SOULOUMIAC, A. (1993). Blind Beamforming for Non-Gaussian Signals. *IEE Proceedings F - Radar and Signal Processing* **140**. https://doi.org/10.1049/ip-f-2.1993.0054

CHEN, A. and BICKEL, P. J. (2006). Efficient independent component analysis. *The Annals of Statistics* **34** 2825 – 2855.

COX, D., LITTLE, J. and OSHEA, D. (2013). *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer Science & Business Media.

DI NARDO, E., GUARINO, G. and SENATO, D. (2009). A new method for fast computing unbiased estimators of cumulants. *Statistics and Computing* **19** 155–165.

EATON, M. L. (2007). *Multivariate statistics*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **53**. A vector space approach, Reprint of the 1983 original [MR0716321].

FISHER, R. A. (1930). Moments and product moments of sampling distributions. *Proceedings of the London Mathematical Society* **2** 199–238.

HANSEN, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica* **50** 1029–1054.

HASTIE, T. and TIBSHIRANI, R. (2002). Independent Components Analysis through Product Density Estimation. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*. *NIPS'02* 665–672. MIT Press, Cambridge, MA, USA.

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

HYVÄRINEN, A., HOYER, P. O. and INKI, M. (2001). Topographic independent component analysis. *Neural computation* **13** 1527–1558.

ILMONEN, P., NORDHAUSEN, K., OJA, H. and OLLILA, E. (2010). A New Performance Index for ICA: Properties, Computation and Asymptotic Analysis. In *Latent Variable Analysis and Signal Separation* (V. VIGNERON, V. ZARZOSO, E. MOREAU, R. GRIBONVAL and E. VINCENT, eds.) 229–236. Springer Berlin Heidelberg, Berlin, Heidelberg.

JAMMALAMADAKA, S. R., TAUFER, E. and TERDIK, G. H. (2021). Asymptotic theory for statistics based on cumulant vectors with applications. *Scandinavian Journal of Statistics* **48** 708-728.

MATTESON, D. S. and TSAY, R. S. (2017). Independent Component Analysis via Distance Covariance. *Journal of the American Statistical Association* **112** 623-637.

MCCULLAGH, P. (2018). *Tensor methods in statistics: Monographs on statistics and applied probability*. Chapman and Hall/CRC.

NEWEY, W. K. and MCFADDEN, D. (1994). Chapter 36 Large sample estimation and hypothesis testing. *Handbook of Econometrics* **4** 2111-2245. Elsevier.

RAO, C. R. (1973). *Linear Statistical Inference and its Applications: Second Editon*. John Wiley & Sons, Inc.

SAMWORTH, R. J. and YUAN, M. (2012). Independent component analysis via nonparametric maximum likelihood estimation. *The Annals of Statistics* **40** 2973 – 3002.

SHAO, X. and ZHANG, J. (2014). Martingale Difference Correlation and Its Use in High-Dimensional Variable Screening. *Journal of the American Statistical Association* **109** 1302–1318.

SHIMIZU, S., HOYER, P. O., HYVÄRINEN, A. and KERMINEN, A. (2006). A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research* **7** 2003–2030.

SPEED, T. P. (1983). Cumulants and partition lattices 1. *Australian Journal of Statistics* **25** 378–388.

SPEED, T. P. (1986). Cumulants and partition lattices II: Generalised k-statistics. *Journal of the Australian Mathematical Society* **40** 34–53.

SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35** 2769 – 2794.

ZWIERNIK, P. (2016). Semialgebraic statistics and latent tree models. *Monographs on Statistics and Applied Probability* **146** 146.