Beyond independent component analysis: identifiability and algorithms

Álvaro Ribot, Anna Seigal, and Piotr Zwiernik

ABSTRACT. Independent Component Analysis (ICA) is a classical method for recovering latent variables with useful identifiability properties. For independent variables, cumulant tensors are diagonal; relaxing independence yields tensors whose zero structure generalizes diagonality. These models have been the subject of recent work in non-independent component analysis. We show that pairwise mean independence answers the question of how much one can relax independence: it is identifiable, any weaker notion is non-identifiable, and it contains the models previously studied as special cases. Our results apply to distributions with the required zero pattern at any cumulant tensor. We propose an algebraic recovery algorithm based on least-squares optimization over the orthogonal group. Simulations highlight robustness: enforcing full independence can harm estimation, while pairwise mean independence enables more stable recovery. These findings extend the classical ICA framework and provide a rigorous basis for blind source separation beyond independence.

1. Introduction

Independent component analysis (ICA) is a tool for blind source separation. It turns linear mixtures into interpretable sources and underpins methods in signal processing, neuroscience, and econometrics [BW97, Car98, MBJS95]. Blind source separation seeks to estimate A and $\mathbf{s}$ from observations of $\mathbf{x} = A\mathbf{s}$, where $\mathbf{x}$ and $\mathbf{s}$ are n-dimensional random vectors, the mixing matrix $A \in \mathbb{R}^{n \times n}$ is fixed and invertible but unknown, and $\mathbf{s}$ is assumed to have mean zero and uncorrelated entries but is otherwise unknown. With no further restrictions on the distribution of $\mathbf{s}$, the solution is not unique. ICA assumes that the entries of $\mathbf{s}$ are independent. Then, the matrix A is unique up to scaling and permutation of columns, provided at most one source is Gaussian [Com94].

In econometrics and other applications, there is growing interest in relaxing the assumption of independent sources [GLS24, HHI01, Jia25, LZS20, Woo95]; see [MZ24] for a detailed discussion. The idea is that full independence of the source variables is too strong. Relaxations of independence offer a more realistic description of real-world systems. However, relax the assumptions too much and $\mathbf{x} = A\mathbf{s}$ is no longer identifiable. The goal is to relax independence while preserving identifiability of A and $\mathbf{s}$ [MZ24, GLS24].

Remark 1.1 (Identifiability). In the component analysis models $\mathbf{x} = A\mathbf{s}$ we consider, rescaling and relabeling the source variables does not affect membership in the model. Hence, at best, we can recover $\mathbf{s}$ up to rescaling and reordering its coordinates, with a corresponding scaling and permutation of the columns of A. We call a model identifiable if we can recover any sufficiently general sources and mixing matrix, up to this relabeling and rescaling.

Sufficiently general, here, means that a certain polynomial in the entries of a higher-order cumulant of $\mathbf{s}$ does not vanish and that the matrix A is invertible.

One popular relaxation replaces independence—the assumption that conditioning gives no information on the distribution—by mean independence, the weaker assumption that conditioning gives no information on the expectation. Concretely, two random variables are *mean independent* when knowing one of them does not affect the expected value of the other, i.e., $\mathbb{E}(x \mid y) = \mathbb{E}(x)$. An illustration comes from the weak form of the Efficient-Market Hypothesis (EMH) [Fam70]: for excess returns r_t and the information set $\mathcal{F}_{t-1}$ known at $t-1$, $\mathbb{E}(r_t \mid \mathcal{F}_{t-1}) = 0$. If a lagged sentiment index s_{t-1} is part of $\mathcal{F}_{t-1}$, then $\mathbb{E}(r_t \mid s_{t-1}) = \mathbb{E}(r_t) = 0$. Related conditions appear in classical measurement error, where the error term u is assumed mean independent of the regressor x ($\mathbb{E}(u \mid x) = 0$) [Woo10], and in randomized controlled trials, where random assignment implies $\mathbb{E}[Y(0) \mid D] = \mathbb{E}[Y(0)]$ [IR15, Chapter 1].

Definition 1.2. For random variables x and y, x is *mean independent* of y if $\mathbb{E}(x \mid y) = \mathbb{E}(x)$. A random vector $\mathbf{x}$ is *pairwise mean independent* if $\mathbb{E}(x_i \mid x_j) = \mathbb{E}(x_i)$ for all $i \neq j$. A random vector $\mathbf{x}$ is *mean independent* if $\mathbb{E}(x_i \mid \mathbf{x}_{\setminus i}) = \mathbb{E}(x_i)$ for all i, where $\mathbf{x}_{\setminus i}$ is the vector $\mathbf{x}$ with coordinate x_i removed.

Independence implies mean independence, mean independence implies pairwise mean independence, and all three imply uncorrelatedness. All notions are distinct. Mean independence and pairwise mean independence appear in [GLS24, Jia25, LZS20, MZ24, Woo95, Woo10], with varying terminology. In [Jia25, GLS24, MZ24] they study mean independence of a random vector and the structure of its moments/cumulants. The papers [Woo95, Woo10] study conditional mean independence (e.g., $\mathbb{E}(x \mid y, z) = \mathbb{E}(x \mid y)$). In [LZS20], mean independence of two random variables is called conditional mean independence.

Our first main result generalizes the main result of [Com94] from independence to pairwise mean independence. It shows the identifiability of component analysis with pairwise mean independent source distributions.

Theorem 1.3 (Identifiability of PMICA). *Consider the model* $\mathbf{x} = A\mathbf{s}$ *where* $A \in \mathbb{R}^{n \times n}$ *is invertible and* $\mathbf{s}$ *is a sufficiently general pairwise mean independent random vector. Then* A *is identifiable from* $\mathbf{x}$ *(up to permutation and scaling of columns).*

Two works lead to Theorem 1.3. First, [MZ24] introduced non-independent component analysis and showed how identifiability can be studied via moment or cumulant tensors. They studied independence via diagonal tensors and obtained first identifiability results for other zero patterns. Second, [RSZ25] showed that a generic symmetric tensor with an orthogonal basis of eigenvectors has a unique such basis (up to signs). The existence of an orthogonal basis of eigenvectors corresponds to a prescribed zero pattern. We show that this zero pattern characterizes pairwise mean independent (PMI) distributions and that sufficiently general PMI distributions have cumulants that are sufficiently general as tensors.

Sufficiently general in Theorem 1.3 means the non-vanishing of a certain polynomial in the entries of a higher-order cumulant of $\mathbf{s}$. A study of these genericity conditions is another main contribution of this paper. In classical ICA, a simple genericity rule guarantees identifiability: at most one Gaussian source. Under PMI the genericity is more subtle: non-Gaussianity does not suffice. We provide low-order, checkable criteria: for $d = 3$, uniqueness holds if and only if at most one third cumulant $\kappa_3(s_i)$ is zero; for $d = 4$, uniqueness holds if and only if the fourth-order cumulants $\kappa_4(s_i)$ are all distinct (Theorem 4.1). For $d \geq 5$ we obtain polynomial non-vanishing conditions. The conditions are illustrated in Figure 1.
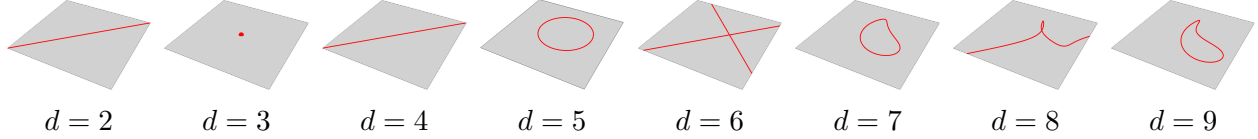
$d = 2 \qquad d = 3 \qquad d = 4 \qquad d = 5 \qquad d = 6 \qquad d = 7 \qquad d = 8 \qquad d = 9$

FIGURE 1. Genericity conditions in the space of d-th order cumulant tensors of pairwise mean independent distributions. The gray plane is the this space, and the red loci are tensors with a non-unique orthogonal basis of eigenvectors. The pictures are three-dimensional slices explained in Remark 4.3.

For an n-dimensional random vector $\mathbf{s}$, the d-th cumulant is a symmetric tensor, an array of format $n \times \cdots \times n$ (d times) whose entries are unchanged under permuting indices. We denote the space of such symmetric tensors by $S^d(\mathbb{R}^n)$ and the dth cumulant of $\mathbf{s}$ by $\kappa_d(\mathbf{s})$. Independence implies diagonal cumulant tensors; see, e.g., [Zwi15]. That is, the cumulants of independent variables $\mathbf{s}$ are a linear subspace of $S^d(\mathbb{R}^n)$ consisting of diagonal tensors

$$(1) \qquad V_{\mathrm{diag}} := \{ \mathcal{T} \in S^d(\mathbb{R}^n) \mid \mathcal{T}_{i_1 \dots i_d} \neq 0 \text{ only if } i_1 = \cdots = i_d \}.$$

There is a corresponding linear subspace for pairwise mean independence:

$$(2) \qquad V_{\mathrm{pmi}} := \{ \mathcal{T} \in S^d(\mathbb{R}^n) \mid \mathcal{T}_{ij \dots j} = 0 \text{ for all } i \neq j \}.$$

For a linear space $V \subset S^d(\mathbb{R}^n)$, we will write $V^{d,n}$ when the ambient space is not clear from the context. We show in Theorem 2.3 that the distributions whose cumulants lie in V_{pmi} for every d are exactly the PMI distributions. When $d = 2$ the two spaces coincide for all n, but they differ for higher d: $\dim V_{\mathrm{diag}} = n$ for all d, whereas $\dim V_{\mathrm{pmi}} = \binom{n+d-1}{d} - 2\binom{n}{2}$ for $d \geq 3$. The set V_{pmi} was introduced in [MZ24, Conjecture 5.17] and used in [RSZ25] to prove generic uniqueness of an orthogonal basis of eigenvectors for tensors. See Table 1 for a summary of our contributions.

Assumption on sources	Cumulants of sources	Generic identifiability
Independence	V_{diag}	yes
Between independence and PMI	V with $V_{\mathrm{diag}} \subseteq V \subseteq V_{\mathrm{pmi}}$	yes
Pairwise mean independence (PMI)	V_{pmi}	yes
Beyond PMI	V with $V_{\mathrm{pmi}} \subsetneq V$	no

TABLE 1. Results at a glance. For $\mathbf{x} = A\mathbf{s}$, A is generically identifiable if $\mathbf{s}$ is PMI (Theorem 1.3) or if its d-th order cumulant tensor lies in V with $V_{\mathrm{diag}} \subseteq V \subseteq V_{\mathrm{pmi}}$ (Theorem 1.4). Identifiability is lost if a mean independence condition is dropped (Theorem 1.5). Linear spaces V_{diag} and V_{pmi} are the cumulant tensors of independent and pairwise mean independent distributions, respectively; see (1) and (2).

Motivated by our identifiability results, we propose an algebraic algorithm to recover pairwise mean independent sources from cumulant tensors. The problem is formulated as a least-squares optimization on the orthogonal group, which we attempt to solve via Riemannian gradient descent (RGD). See Algorithm 1.

Algorithm 1 `RGD-PMICA`: Finding PMI sources

Input: Data matrix $X \quad \in \mathbb{R}^{N \times n}$ $\qquad\qquad$ ▷ $N = \#$ samples, $n = \#$ features
$\quad X_w \leftarrow \texttt{Whitening}(X) \quad \in \mathbb{R}^{N \times n}$ $\qquad\qquad$ ▷ Center the mean and decorrelate
$\quad \widehat{\kappa} \leftarrow \texttt{CumulantTensor4}(X_w) \quad \in S^4(\mathbb{R}^n)$ $\qquad$ ▷ Estimate fourth-order cumulant tensor
$\quad \widehat{Q} \leftarrow \arg\min_{Q \in O(n)} \sum_{i \neq j} [Q^\top \bullet \widehat{\kappa}]_{ijjj}^2 \quad \in \mathbb{R}^{n \times n}$ $\quad$ ▷ Minimize distance to V_{pmi} with RGD
Output: $S \leftarrow X_w \widehat{Q} \quad \in \mathbb{R}^{N \times n}$ $\qquad\qquad\qquad\qquad$ ▷ PMI sources

---

A practical route to identifiability is to impose a *generative* structure that forces certain entries of $\kappa_d(\mathbf{s})$ to vanish. Popular examples are *topographic ICA* [HHI01] and, more generally, *correlated-energy* models $s_i = \sigma_i \varepsilon_i$ with independent ε_i and exogenous scales σ_i, which imply PMI and hence $\kappa_d(\mathbf{s}) \in V_{\text{pmi}}$ (see Section 5.1.5).

The next result shows that any such family inherits generic identifiability from a single d-th order cumulant once $\kappa_d(\mathbf{s})$ is sufficiently general in the corresponding linear space. As in Theorem 1.3, a source distribution $\mathbf{s}$ is sufficiently general if the sources $\mathbf{s} = (s_1, \ldots, s_n)$ have at most one third cumulant $\kappa_3(s_i)$ equal to zero, or if all fourth order cumulants $\kappa_4(s_i)$ are distinct.

Theorem 1.4. *Consider the model* $\mathbf{x} = A\mathbf{s}$ *where* $A \in \mathbb{R}^{n \times n}$ *is invertible and* $\mathbf{s}$ *is a random vector with identity covariance matrix such that, for some* $d \geq 3$, $\kappa_d(\mathbf{s})$ *is sufficiently general in* V, *where* $V_{\text{diag}} \subseteq V \subseteq V_{\text{pmi}}$. *Then,* A *is identifiable from* $\kappa_d(\mathbf{x})$ *(up to permutation and sign-flip of columns).*

We also show that PMI is maximal for identifiability: models are generically identifiable under pairwise mean independence, but generically unidentifiable once any such assumption is dropped. Each mean independence assumption imposes one zero restriction on each cumulant tensor.

Theorem 1.5. *Consider the model* $\mathbf{x} = A\mathbf{s}$ *where* $A \in \mathbb{R}^{n \times n}$ *is invertible and* s_i *is mean independent of* s_j *for all pairs* $i \neq j$ *except one. Then* A *is generically unidentifiable from* $\kappa_d(\mathbf{x})$ *for any* $d \geq 3$ *(up to permutation and scaling of columns).*

The rest of this paper is organized as follows. We characterize the cumulants of pairwise mean independent distributions in Section 2. We prove Theorem 1.3 in Section 3. We prove Theorem 1.4 and describe what it means for a distribution or cumulant tensor to be sufficiently general in Section 4. We prove Theorem 1.5 and discuss other relaxations of independence under which identifiability holds in Section 5. We investigate the consistency and sample complexity of our minimum-distance estimator in Section 6. We analyze when ICA methods recover PMI distributions in Section 7. Our numerical experiments are in Section 8.

2. Cumulants of pairwise mean independent distributions

This section links pairwise mean independence to the linear space V_{pmi} in (2). Given an n-dimensional random vector $\mathbf{z}$, let $M_{\mathbf{z}}(\mathbf{t}) = \mathbb{E}(e^{\mathbf{t}^\top \mathbf{z}})$ and $K_{\mathbf{z}}(\mathbf{t}) = \log M_{\mathbf{z}}(\mathbf{t})$ be the moment-generating function and cumulant-generating function of $\mathbf{z}$, respectively. The d-th moment tensor is $\mu_d(\mathbf{z}) = \mathbb{E}(\mathbf{z}^{\otimes d})$. Its entries are the d-th order partial derivatives of $M_{\mathbf{z}}(\mathbf{t})$ evaluated at $\mathbf{t} = 0$. The d-th cumulant tensor $\kappa_d(\mathbf{z})$ has entries $\kappa_d(\mathbf{z})_{i_1,\ldots,i_d} = \frac{\partial^d}{\partial t_{i_1} \cdots \partial t_{i_d}} K_{\mathbf{z}}(\mathbf{t})\big|_{\mathbf{t}=0}$, whenever the corresponding partial derivative exists.

The second moment and cumulant are symmetric matrices. The higher-order moments and cumulants are symmetric tensors: a tensor $\mathcal{T}$ is symmetric if $\mathcal{T}_{i_1,\ldots,i_d} = \mathcal{T}_{i_{\sigma(1)},\ldots,i_{\sigma(d)}}$ for any $i_1,\ldots,i_d \in [n]$ and any permutation $\sigma \in S_d$. If $d = 2$, this recovers the definition of symmetric matrices. Let $S^d(\mathbb{R}^n)$ be the set of real symmetric $n \times \cdots \times n$ tensors of order d.

We write $V_{\text{pmi}}^{d,n}$ instead of V_{pmi} to make the tensor format explicit: $V_{\text{pmi}}^{d,n} \subset S^d(\mathbb{R}^n)$. For $d = 2$, $V_{\text{pmi}}^{2,n} = V_{\text{diag}}^{2,n}$ is the space of diagonal $n \times n$ matrices.

Theorem 2.1. *If $\mathbf{z} = (z_1,\ldots,z_n)$ has independent coordinates, then $\kappa_d(\mathbf{z}) \in V_{\text{diag}}^{d,n}$ for all $d \geq 2$ for which the corresponding moments of order d exist. Conversely, if all off-diagonal entries of $\kappa_d(\mathbf{z})$ vanish for every $d \geq 2$ and the cumulant-generating function $K_{\mathbf{z}}(\mathbf{t})$ is finite in a neighborhood of 0, then the coordinates of $\mathbf{z}$ are independent.*

Remark 2.2. The forward implication is classical (vanishing mixed cumulants under independence). For the converse, finiteness of $K_{\mathbf{z}}$ near 0 implies that $K_{\mathbf{z}}(\mathbf{t}) = \sum_{i=1}^{n} K_{z_i}(t_i)$, hence $M_{\mathbf{z}}(\mathbf{t}) = \prod_{i=1}^{n} M_{z_i}(t_i)$ in a neighborhood of 0, which characterizes independence.

The next result shows that V_{pmi}, defined in (2), plays for pairwise mean independence the role that V_{diag} plays for independence. One direction was given in [GLS24, Proposition 3.3]. We give a detailed explanation and also show the converse direction.

Theorem 2.3. *If $\mathbf{z} = (z_1,\ldots,z_n)$ is pairwise mean independent then $\kappa_d(\mathbf{z}) \in V_{\text{pmi}}^{d,n}$ for all $d \geq 2$ for which the corresponding moments of order d exist. Conversely, if $\kappa_d(\mathbf{z}) \in V_{\text{pmi}}^{d,n}$ for all $d \geq 2$ and $K_{\mathbf{z}}(\mathbf{t})$ is finite in a neighborhood of 0, then $\mathbf{z}$ is pairwise mean independent.*

PROOF. Fix $d \geq 2$ such that all moments of order d exist. Then all moments and cumulants up to order d exist. The $(i,j,\ldots,j)$ entry of the d-th order moments tensor $\mu_d(\mathbf{z})$ equals $\mathbb{E}(z_i z_j^{d-1})$. By the tower property,

$$(3) \qquad \mathbb{E}(z_i z_j^{d-1}) = \mathbb{E}\big(z_j^{d-1}\,\mathbb{E}(z_i \mid z_j)\big) = \mathbb{E}(z_i)\mathbb{E}(z_j^{d-1}) \qquad (i \neq j).$$

Let Π_d be the partition lattice of $\{1,\ldots,d\}$ and, for $\pi \in \Pi_d$, denote by $|\pi|$ the number of blocks of π. The cumulant-moment relation is

$$(4) \qquad \kappa_d(\mathbf{z})_{i_1,\ldots,i_d} = \sum_{\pi \in \Pi_d} (-1)^{|\pi|-1}(|\pi|-1)! \prod_{B \in \pi} \mathbb{E}\Big(\prod_{k \in B} z_{i_k}\Big),$$

see, for example, [Zwi15, §4.2.1]. Specialize to $(i_1,\ldots,i_d) = (i,j,\ldots,j)$ with $i \neq j$ and let $\pi_0 := 1\,|\,2\cdots d$ be a partition with two blocks $\{1\}$ and $\{2,\ldots,d\}$. For any $\pi \in \Pi_d$, write $\pi \wedge \pi_0$ for the partition obtained by splitting the block of π that contains 1 into $\{1\}$ and the rest. By (3),

$$\prod_{B \in \pi} \mathbb{E}\Big(\prod_{k \in B} z_{i_k}\Big) = \prod_{B \in \pi \wedge \pi_0} \mathbb{E}\Big(\prod_{k \in B} z_{i_k}\Big).$$

Hence

$$\kappa_d(\mathbf{z})_{i,j,\ldots,j} = \sum_{\pi \in \Pi_d} (-1)^{|\pi|-1}(|\pi|-1)! \prod_{B \in \pi \wedge \pi_0} \mathbb{E}\Big(\prod_{k \in B} z_{i_k}\Big)$$

$$= \sum_{\delta \leq \pi_0} \Bigg(\sum_{\pi:\, \pi \wedge \pi_0 = \delta} (-1)^{|\pi|-1}(|\pi|-1)! \Bigg) \prod_{B \in \delta} \mathbb{E}\Big(\prod_{k \in B} z_{i_k}\Big),$$

and the inner sum vanishes by Möbius inversion on Π_d [Zwi15, Lemma 4.19], so $\kappa_d(\mathbf{z}) \in V_{\text{pmi}}^{d,n}$.

For the converse, we use a standard L^2 projection argument. Here L^2 denotes the set of all square integrable functions with respect to the fixed probability space and with the standard inner product. Cumulants are invariant under mean shifts, so assume $\mathbb{E}z_i = 0$. Define the Hilbert space of square-integrable functions of W by

$$\mathcal{H}(W) := \{\, h(W) : \ h \text{ measurable and } \mathbb{E}[h(W)^2] < \infty \,\}.$$

Here, $\mathcal{H}(W) \subseteq L^2$ and the conditional expectation $\mathbb{E}(\cdot|W)$ gives the orthogonal projection from L^2 to $\mathcal{H}(W)$; see, e.g., Exercise 34.12 in [Bil95]. Let $g(z_j) := \mathbb{E}(z_i \mid z_j) \in \mathcal{H}(z_j)$. The vanishing of all mixed entries $\kappa_{k+1}(\mathbf{z})_{i,j,\dots,j} = 0$ for $k \geq 1$ implies (by (4) as above) that

$$\mathbb{E}\big(g(z_j)\, z_j^k\big) = 0 \qquad \text{for all } k \geq 0.$$

Since $K_{\mathbf{z}}$ is finite near 0, $\{z_j^k : k \geq 0\}$ is dense in $\mathcal{H}(z_j)$ (polynomial density; see [Akh20, Cor. 2.3.3]). Hence $g(z_j) \in \mathcal{H}(z_j)$ is orthogonal to a dense subset of $\mathcal{H}(z_j)$, so $g(z_j) = 0$ in L^2, i.e., $\mathbb{E}(z_i \mid z_j) = 0$ almost surely. This holds for all $i \neq j$, proving pairwise mean independence. $\qquad\square$

Corollary 2.4. *Theorem 2.3 remains true with μ_d in place of κ_d provided $\mathbb{E}(\mathbf{z}) = 0$.*

3. Pairwise Mean Independent Component Analysis

We prove Theorem 1.3; i.e., we show that we can identify A in the model $\mathbf{x} = A\mathbf{s}$ if $\mathbf{s}$ is pairwise mean independent and sufficiently general. This gives the identifiability of Pairwise Mean Independent Component Analysis (PMICA). Our approach is to relate the identifiability of PMICA to the study of orthogonal eigenvectors of tensors [RSZ25].

We first explain how to reduce from an invertible matrix to an orthogonal matrix in our component analysis model, a standard procedure that sometimes goes by the name of whitening. Given a random vector $\mathbf{y}$, we define its transformation $\mathbf{y}_w$ to be $Q\Lambda^{-1/2}Q^\top(\mathbf{y} - \mathbb{E}(\mathbf{y}))$ where $Q\Lambda Q^\top$ is the eigendecomposition of the covariance matrix $\mathrm{Cov}(\mathbf{y}) = \mathbb{E}((\mathbf{y} - \mathbb{E}(\mathbf{y}))(\mathbf{y} - \mathbb{E}(\mathbf{y}))^\top)$. Then $\mathbb{E}(\mathbf{y}_w) = 0$ and $\mathbb{E}(\mathbf{y}_w\mathbf{y}_w^\top) = I$. This shifts and rescales the variables so that they have mean zero and variance one. We can assume, without loss of generality, that $\mathrm{Cov}(\mathbf{s}) = I$, as follows. If $\mathbf{s}$ is PMI, then it has uncorrelated entries, and scaling $\mathbf{s}$ does not change membership in the PMICA model because V_{pmi} is given by zero restrictions and the scalars can be absorbed by the mixing matrix A. Applying the procedure explained above to random variables $\mathbf{x}$ in a component analysis model $\mathbf{x} = A\mathbf{s}$ turns the equation into $\mathbf{x}_w = \tilde{A}\mathbf{s}$, where $\tilde{A}$ is now an orthogonal matrix, since $I = \mathrm{Cov}(\mathbf{x}_w) = \tilde{A}\,\mathrm{Cov}(\mathbf{s})\tilde{A}^\top = \tilde{A}\tilde{A}^\top$.

The orthogonal group is $\mathrm{O}(n) = \{A \in \mathbb{R}^{n \times n} \mid A^\top A = I\}$. It acts on $S^d(\mathbb{R}^n)$ as follows: given $A \in \mathrm{O}(n)$ and $\mathcal{T} \in S^d(\mathbb{R}^n)$, define $A \bullet \mathcal{T} \in S^d(\mathbb{R}^n)$ by

$$[A \bullet \mathcal{T}]_{i_1,\dots,i_d} := \sum_{j_1,\dots,j_d=1}^{n} A_{i_1 j_1} \cdots A_{i_d j_d} \mathcal{T}_{j_1 \dots j_d}.$$

Given a linear space $V \subseteq S^d(\mathbb{R}^n)$, its orbit is $\mathrm{O}(n) \bullet V = \{A \bullet \mathcal{T} \mid A \in \mathrm{O}(n), \mathcal{T} \in V\}$. The following result is well-known.

Lemma 3.1. *Moment and cumulant tensors are multilinear: For every $d \geq 2$ and every $A \in \mathbb{R}^{n \times n}$ it holds that $\mu_d(A\mathbf{x}) = A \bullet \mu_d(\mathbf{x})$, $\kappa_d(A\mathbf{x}) = A \bullet \kappa_d(\mathbf{x})$.*

Let $\mathbf{x} = A\mathbf{s}$, where $\mathbf{s}$ are (fully) independent with mean zero and unit variance, and let $A \in \mathrm{O}(n)$. The d-th cumulant of $\mathbf{x}$ has the form

$$(5) \qquad \kappa_d(\mathbf{x}) = A \bullet \kappa_d(\mathbf{s}) = \sum_{j=1}^{n} \kappa_d(s_j)\mathbf{a}_j^{\otimes d},$$

where $\mathbf{a}_j$ is the j-th column of $A \in \mathrm{O}(n)$ and $\kappa_d(s_j) = \kappa_d(\mathbf{s})_{j,\ldots,j}$ is the d-th cumulant of source variable s_j. This writes the cumulant as a sum of outer products of orthogonal vectors, so $\kappa_d(\mathbf{x})$ is an orthogonally decomposable (odeco) tensor.

The cumulant expression (5) relates usual ICA to orthogonal tensor decomposition. Odeco decompositions are unique [AGH+14, Theorem 4.1], so the matrix is identifiable (up to sign and permutation) from $\kappa_d(\mathbf{x})$ if and only if at most one $\kappa(s_j)$ is zero. The classical identifiability result for ICA [Com94] says that identifiability holds if and only if at most one source is Gaussian. This relates to the tensor decomposition as follows: the Gaussian distribution has zero d-th cumulants for all $d \geq 3$ and it is the only probability distribution with the property that there exists d_0 such that the d-th cumulant vanishes for all $d \geq d_0$ [Mar39]. Hence, for non-Gaussian sources, we can carry out tensor decomposition of higher-order cumulants to recover A, see [LCMS25, Section 2.1].

We have seen that orthogonal decompositions of symmetric tensors play a role in independent component analysis. In ICA, the decompositions have core tensors in V_{diag}. In this section, we study decompositions with core tensors in V_{pmi} as defined in (2). Given $\mathcal{T} \in S^d(\mathbb{R}^n)$, a unit norm vector $v \in \mathbb{R}^n$ is an *eigenvector* of $\mathcal{T}$ with *eigenvalue* $\lambda \in \mathbb{R}$ if $\mathcal{T}(\cdot, v, \ldots, v) = \lambda v$, viewing $\mathcal{T}$ as a multilinear map. We build upon the following result.

Theorem 3.2 ([RSZ25, Theorem 1.1]). *If $\mathcal{T} \in S^d(\mathbb{R}^n)$ is a generic symmetric tensor with an orthogonal basis of eigenvectors, then this basis is unique (up to sign flip).*

This result is key to our discussion for the following reason. Let $\mathrm{SP}(n) \subset \mathrm{O}(n)$ denote the set of $n \times n$ signed permutation matrices, i.e., matrices of the form DP where D is diagonal with diagonal entries ± 1 and P is a permutation matrix. The set of symmetric tensors in $S^d(\mathbb{R}^n)$ with an orthogonal basis of eigenvectors is the orbit $\mathrm{O}(n) \bullet V_{\mathrm{pmi}}$. This fact, together with Theorem 3.2, implies that if $\mathcal{T} \in V_{\mathrm{pmi}} \subset S^d(\mathbb{R}^n)$ is generic, then $Q \bullet \mathcal{T} \in V_{\mathrm{pmi}}$ if and only if $Q \in \mathrm{SP}(n)$ (see Propositions 4.2 and 6.1 in [RSZ25] for details).

The following result implies that general independent distributions are sufficiently general as pairwise mean independent distributions, which we use later to prove Theorem 1.3.

Lemma 3.3. *A generic symmetric odeco tensor in $S^d(\mathbb{R}^n)$ has a unique orthogonal basis of eigenvectors (up to sign flip).*

PROOF. Consider an odeco tensor $\mathcal{T} = \sum_{j=1}^{n} \lambda_j q_j^{\otimes d} \in S^d(\mathbb{R}^n)$ where $\lambda_1, \ldots, \lambda_n \neq 0$ and $\{q_1, \ldots, q_n\}$ is an orthonormal basis of $\mathbb{R}^n$. Then $\mathcal{T}$ has $\frac{(d-1)^n - 1}{d-2}$ different eigenvectors (up to scaling) in $\mathbb{C}^n$ given as follows: for any $1 \leq k \leq n$, any $\mathcal{J} = \{j_1, \ldots, j_k\} \subseteq [n]$, and any $(k-1)$-tuple $\eta_1, \ldots, \eta_{k-1}$ of $(d-2)$-nd roots of unity, there is one eigenvector $v \in \mathbb{C}^n$ whose coordinates with respect to the basis $\{q_1, \ldots, q_n\}$ are

$$v_j := \langle v, q_j \rangle = \begin{cases} \eta_l \lambda_{j_l}^{-\frac{1}{d-2}} & \text{if } j = j_l \text{ for some } l \in \{1, \ldots, k-1\} \\ \lambda_{j_k}^{-\frac{1}{d-2}} & \text{if } j = j_k \\ 0 & \text{if } j \notin \mathcal{J} \end{cases}$$

(see [Rob16, Theorem 2.3]). Consider a collection $\mathcal{J}_1, \ldots, \mathcal{J}_n$ of non-empty subsets of $[n]$, they are pairwise disjoint if and only if $\{\mathcal{J}_1, \ldots, \mathcal{J}_n\} = \{\{1\}, \ldots, \{n\}\}$, which corresponds to the set of eigenvectors $\{q_1, \ldots, q_n\}$. Suppose that $\{\mathcal{J}_1, \ldots, \mathcal{J}_n\} \neq \{\{1\}, \ldots, \{n\}\}$ and, without loss of generality, suppose that $\mathcal{J}_1 \cap \mathcal{J}_2 \neq \varnothing$. For any $(k-1)$-tuple of $(d-2)$-nd roots of unity, imposing orthogonality between the eigenvectors corresponding to $\mathcal{J}_1$ and $\mathcal{J}_2$ leads to a non-zero polynomial expression with complex coefficients in the parameters $\left\{ \lambda_j^{-\frac{1}{d-2}} \mid j \in \mathcal{J}_1 \cap \mathcal{J}_2 \right\}$. This polynomial is non-zero for generic $\lambda_1, \ldots, \lambda_n$. $\qquad\square$

Remark 3.4. The previous result shows that the orthogonal basis of eigenvectors of a general odeco tensor is unique among all complex eigenvectors. The genericity of an odeco tensor $\mathcal{T} = \sum_{i=1}^n \lambda_i q_i^{\otimes d} \in S^d(\mathbb{R}^n)$ depends only on the scalars $\lambda_1, \ldots, \lambda_n$, not on the orthonormal basis $\{q_1, \ldots, q_n\}$ (see Theorem 4.4).

PROOF OF THEOREM 1.3. The random vector $\mathbf{s}$ is mean independent, so it has uncorrelated entries. Rescaling its entries does not change membership of $\kappa_d(\mathbf{s}) \in V_{\mathrm{pmi}}$, since V_{pmi} is given by zero restrictions. Therefore, we can assume that $\mathrm{Cov}(\mathbf{s}) = I$ without loss of generality. We can also assume that $A \in \mathrm{O}(n)$, after whitening $\mathbf{x}$. We show that A can be recovered from the d-th order cumulant tensor $\kappa_d(\mathbf{x})$ for some $d \geq 3$ provided $\mathbf{s}$ is sufficiently general. The idea is to apply Theorem 3.2 to $\kappa_d(\mathbf{x})$: from Theorem 2.3 we know that $\kappa_d(\mathbf{x}) \in \mathrm{O}(n) \bullet V_{\mathrm{pmi}}$, and using Lemma 3.3 we can ensure that $\kappa_d(\mathbf{x})$ is sufficiently general. Consider a random vector $\mathbf{s}$ with n pairwise mean independent entries. The d-th order cumulant tensor of $\mathbf{s}$ lies in V_{pmi}, by Theorem 2.3. Consider a random vector $\mathbf{s}^{(0)}$ with n independent entries that is independent from $\mathbf{s}$ and such that the corresponding d-th cumulant (odeco) tensor $\kappa_d(\mathbf{s}^{(0)}) = \sum_{i=1}^n \kappa_d(s_i^{(0)}) e_i^{\otimes d}$ is generic in the sense of Lemma 3.3. For each $\alpha \in \mathbb{R}$, let $\mathbf{s}^{(\alpha)} = \alpha \mathbf{s} + (1-\alpha)\mathbf{s}^{(0)}$. Since $\mathbf{s}$ and $\mathbf{s}^{(0)}$ are independent and by multilinearity of cumulants, we have $\kappa_d(\mathbf{s}^{(\alpha)}) = \alpha^d \kappa_d(\mathbf{s}) + (1-\alpha)^d \kappa_d(\mathbf{s}^{(0)})$ and so $\kappa_d(\mathbf{s}^{(\alpha)}) \in V_{\mathrm{pmi}}^d$. Therefore, for a generic $\alpha \in \mathbb{R}$, $\kappa_d(\mathbf{s}^{(\alpha)})$ is a sufficiently generic tensor of V_{pmi}^d. Hence, if $\mathbf{s} = \mathbf{s}^{(1)}$ is a sufficiently general distribution, the mixing matrix A can be recovered from $\kappa_d(\mathbf{x})$ up to right-multiplication by $\mathrm{SP}(n)$, by Theorem 3.2. $\qquad\square$

Theorem 1.3 pertains to sufficiently general distributions. Our proof requires that, for some $d \geq 3$, the d-th order cumulant of $\mathbf{s}$ is generic in V_{pmi}, meaning it has a unique orthogonal basis of eigenvectors. This genericity condition may be relaxed if we consider multiple higher-order cumulants. Characterizing sufficiently general PMI distributions assuming access to all higher-order cumulants is a direction for future work. In the next section, we discuss genericity conditions fixing the order of the cumulant.

A relevant comparison is ICA. There, a sufficiently general distribution is one with at most one Gaussian source. However, non-Gaussianity is not sufficient to recover the sources from a fixed d-th order cumulant: one also needs the d-th cumulant of each non-Gaussian source to be nonzero. In principle, one may need many cumulants to recover the sources. However, non-Gaussianity is not sufficient for PMICA. For example, if $\mathbf{z}$ is spherical (i.e. $\mathbf{z}$ and $Q\mathbf{z}$ have the same distribution for all orthogonal matrices Q) then $\mathbf{z}$ is pairwise mean independent but the cumulants contain no information about the rotation Q. Being non-spherical is also not sufficient: rotating the uniform distribution on $[-1,1]^2$ by an angle of $\pi/4$ gives a PMI distribution.

4. Sufficiently general moments and cumulants

So far we have studied identifiability for sufficiently general tensors or distributions. Now we investigate what it means to be sufficiently general.

Theorem 4.1. *Let $\mathcal{T} \in V_{\mathrm{pmi}} \subset S^d(\mathbb{R}^n)$. Then $\mathcal{T}$ has a unique orthogonal basis of eigenvectors if and only if*
- *If $d = 2$, its diagonal entries are distinct.*
- *If $d = 3$, at most one of its diagonal entries is zero.*
- *If $d = 4$, its diagonal entries are distinct.*

PROOF. For $d = 2$, the statement follows from the spectral theorem. Let $d \geq 3$, we study the binary case first, i.e., let $n = 2$ and $V_{\mathrm{pmi}} \subset S^d(\mathbb{R}^2)$. We denote the canonical basis of $\mathbb{R}^2$ by $\{e_0, e_1\}$. This way, the coordinates of tensors in $S^d(\mathbb{R}^2)$ are specified by binary strings of length d. For a symmetric tensor $\mathcal{T} \in S^d(\mathbb{R}^n)$, its entry $\mathcal{T}_\mathbf{i}$ is determined by $|\mathbf{i}| = \sum_{k=1}^d i_k$. Therefore, a symmetric tensor in $S^d(\mathbb{R}^2)$ is specified by $d+1$ parameters $t_0, \ldots, t_d$ where $\mathcal{T}_\mathbf{i} = t_{|\mathbf{i}|}$ for all $\mathbf{i} \in \{0, 1\}^d$. This way, we have

$$V_{\mathrm{pmi}} = \{\mathcal{T} \in S^d(\mathbb{R}^2) \mid t_1 = t_{d-1} = 0\}$$

and its orthogonal complement is

$$V_{\mathrm{pmi}}^\perp = \{\mathcal{T} \in S^d(\mathbb{R}^2) \mid t_0 = t_2 = \cdots = t_{d-2} = t_d = 0\}.$$

Given $Q \in \mathrm{O}(2)$, the linear map $Q\bullet : S^d(\mathbb{R}^2) \to S^d(\mathbb{R}^2), \mathcal{T} \mapsto Q \bullet \mathcal{T}$ can be represented by a $(d+1) \times (d+1)$ matrix indexed by $\{0, 1, \ldots, d\}^2$. We define M_Q as the $2 \times (d-1)$ submatrix whose rows are indexed by $\{1, d-1\}$ and columns are indexed by $\{0, 2, \ldots, d-2, d\}$. The k-th column of this matrix is

$$[M_Q]_k = \begin{pmatrix} \binom{d-1}{k} q_{00}^{d-k-1} q_{01}^k q_{10} + \binom{d-1}{k-1} q_{00}^{d-k} q_{01}^{k-1} q_{11} \\ \binom{d-1}{k} q_{00} q_{10}^{d-k-1} q_{11}^k + \binom{d-1}{k-1} q_{01} q_{10}^{d-k} q_{11}^{k-1} \end{pmatrix}.$$

Given $\mathcal{T} \in V_{\mathrm{pmi}} \subset S^d(\mathbb{R}^2)$, we have $Q \bullet \mathcal{T} \in V_{\mathrm{pmi}}$ if and only if $(t_0, t_2, \ldots, t_{d-2}, t_d) \in \ker M_Q$. Note that $M_Q = 0$ if and only if $Q \in \mathrm{SP}(2)$. We are interested in $Q \notin \mathrm{SP}(2)$. Let us assume, without loss of generality, that $Q \in \mathrm{SO}(2)$. That is,

$$Q = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \in \mathrm{SO}(2),$$

with $a^2 + b^2 = 1$ and $a, b, \neq 0$. Indeed, if $Q \in \mathcal{T} \in V_{\mathrm{pmi}}$ for some $Q \in \mathrm{O}(2) \setminus \mathrm{SO}(2)$ and $\tilde{Q} \in \mathrm{SO}(2)$ is obtained by flipping a column from Q, then $\tilde{Q} \bullet \mathcal{T} \in V_{\mathrm{pmi}}$. If $d = 3$, then

$$M_Q = ab \begin{pmatrix} a & b \\ b & -a \end{pmatrix} \implies \mathrm{rank}(M_Q) = 2 \text{ for all } a, b \neq 0.$$

Therefore, $\mathcal{T} \in V_{\mathrm{pmi}} \subset S^3(\mathbb{R}^2)$ has a unique orthogonal basis of eigenvectors if and only if $(t_0, t_3) \neq (0, 0)$, i.e., $\mathcal{T} \neq 0$. If $d = 4$, then

$$M_Q = ab \begin{pmatrix} a^2 & -3(a^2 - b^2) & -b^2 \\ b^2 & 3(a^2 - b^2) & -a^2 \end{pmatrix} \implies \ker(M_Q) = \begin{cases} \langle (3, 1, 3) \rangle & \text{if } a \neq \pm b \\ \langle (1, 0, 1), (0, 1, 0) \rangle & \text{if } a = \pm b. \end{cases}$$

So $\mathcal{T} \in V_{\mathrm{pmi}} \subset S^4(\mathbb{R}^2)$ has a unique orthogonal basis of eigenvectors if and only if $t_0 \neq t_4$.

Next, we reduce the case of $n > 2$ to the binary case. Given $Q \in \mathrm{O}(n)$, its normal form is a decomposition $Q = P^\top R P$, where $P \in \mathrm{O}(n)$ and R is block diagonal with 2×2 blocks

in $O(2)$; see, e.g., [Rom05, Theorem 10.19]. We have $\mathcal{T} \in V_{\mathrm{pmi}}$ and $Q \bullet \mathcal{T} \in V_{\mathrm{pmi}}$ if and only if $\mathcal{S} := P \bullet \mathcal{T} \in P \bullet V_{\mathrm{pmi}}$ and $R \bullet \mathcal{S} \in P \bullet V_{\mathrm{pmi}}$. Moreover, the coordinates of $\mathcal{S}$ and $R \bullet \mathcal{S}$ in the basis $\{p_{i_1} \otimes \cdots \otimes p_{i_d} \mid i_k \in [n]\}$ are $\mathcal{T}_{i_1,\ldots,i_d}$ and $[Q \bullet \mathcal{T}]_{i_1,\ldots,i_d}$, respectively, where p_j is the j-th column of P; see [RSZ25, Lemma 5.12]. This reduces the problem to the binary case studied above. Therefore, $\mathcal{T} \in V_{\mathrm{pmi}} \subset S^3(\mathbb{R}^n)$ has a unique basis of orthogonal eigenvectors if and only if every pair of its diagonal entries are not simultaneously zero, i.e., it has at most one zero diagonal entry. For $d = 4$, $\mathcal{T} \in V_{\mathrm{pmi}} \subset S^4(\mathbb{R}^n)$ has a unique basis of orthogonal eigenvectors if and only if every pair of its diagonal entries are distinct. $\square$

Remark 4.2. For PMICA, Theorem 4.1 implies that PMI sources must follow different distributions if one wishes to identify them from the fourth-order cumulant/moment tensors.

In Theorem 4.1 we focus on the cases $d \le 4$ because these order moments/cumulants are most commonly studied in ICA and its related methods. However, the same proof can be extended to find the genericity conditions for higher d. We include the next few cases in the following remark. It is an open problem to resolve the genericity conditions for all d. Following the trend below, we might expect to obtain $\binom{n}{2}$ irreducible polynomials of degree $d - 3$ for $d \ge 7$.

Remark 4.3. Figure 1 shows the genericity conditions in $V_{\mathrm{pmi}} \subset S^d(\mathbb{R}^2)$ for $2 \le d \le 9$. Those pictures are of the three-dimensional slice obtained by setting $t_1 = 0$ and $t_k = k$ for $1 < k < d - 1$, following the notation introduced in the proof above. The cases where $d \le 4$ were addressed in Theorem 4.1. For $d \ge 5$, having distinct or nonzero diagonal entries is not sufficient for $\mathcal{T} \in V_{\mathrm{pmi}} \subset S^d(\mathbb{R}^n)$ to have a unique basis of orthogonal eigenvectors, since the kernel of the matrix M_Q studied in the proof of Theorem 4.1 is more complicated. For example, for $d = 5$ we get

$$\ker(M_Q) = \langle (4(a^2 - b^2), a^2 - b^2, -ab, 2ab), (-2ab, ab, a^2 - b^2, 4(a^2 - b^2)) \rangle,$$

so $\mathcal{T} \in V_{\mathrm{pmi}} \subset S^5(\mathbb{R}^2)$ has a unique basis of orthogonal eigenvectors if and only if

$$t_0^2 - 2\,t_0 t_2 - 8\,t_2^2 - 8\,t_3^2 - 2\,t_3 t_5 + t_5^2 \ne 0.$$

For $d = 6$, we get that $\ker(M_Q)$ is

$$\begin{cases} \langle (5,1,0,1,5), (5(a^2 - b^2), a^2 - b^2, -ab, 0, 0), (-10ab, 0, a^2 - b^2, -2ab, 0) \rangle & \text{if } a \ne \pm b \\ \langle (-5,1,0,0,0), (0,0,1,0,0), (5,0,0,1,0), (1,0,0,0,1) \rangle & \text{if } a = \pm b \end{cases}$$

so $\mathcal{T} \in V_{\mathrm{pmi}} \subset S^6(\mathbb{R}^2)$ has a unique basis of orthogonal eigenvectors if and only if

$$t_0 - 5\,t_2 - 5\,t_4 + t_6 \ne 0 \quad \text{and} \quad t_0 + 5\,t_2 - 5\,t_4 - t_6 \ne 0.$$

For $d = 7$, the genericity conditions in V_{pmi} are given by a quartic with 63 monomials. For $d = 8$, they are given by a quintic with 204 monomials. For $d = 9$, they are given by a sextic with 752 monomials[1]. Using the same argument as above, the genericity conditions in $S^d(\mathbb{R}^n)$ can be obtained by reducing the problem to the binary case and arguing about pairs of coordinates. For example, for $d = 5$ we obtain that $\mathcal{T} \in V_{\mathrm{pmi}} \subset S^5(\mathbb{R}^n)$ has a unique basis of orthogonal eigenvectors if and only if

$$\mathcal{T}_{iiiii}^2 - 2\mathcal{T}_{iiiii}\mathcal{T}_{iiijj} - 8\mathcal{T}_{iiijj}^2 - 8\mathcal{T}_{iijjj}^2 - 2\mathcal{T}_{iijjj}\mathcal{T}_{jjjjj} + \mathcal{T}_{jjjjj}^2 \ne 0 \quad \text{for all } i \ne j.$$

---

[1]We did these computations using `Macaulay2` [GS].

Recall that $V_{\text{diag}} \subseteq V_{\text{pmi}}$ and that the odeco tensors are the ones that lie in V_{diag} after an orthogonal change of basis. In Lemma 3.3 we showed that a generic symmetric odeco tensor has a unique orthogonal basis of eigenvectors. The following result specifies what generic means in this context.

Theorem 4.4. *Let $\mathcal{T} \in V_{\text{diag}} \subset S^d(\mathbb{R}^n)$.*

- *If d is even, $\mathcal{T}$ has a unique basis of orthogonal eigenvectors if and only if its diagonal entries are distinct.*
- *If d is odd, $\mathcal{T}$ has a unique basis of orthogonal eigenvectors if and only if at most one of its diagonal entries is zero.*

PROOF. We use the same proof idea as in Theorem 4.1. First, let $n = 2$. Let $\tilde{M}_Q$ be the 2×2 submatrix of the matrix M_Q defined in the proof of Theorem 4.1 corresponding to the columns indexed by 0 and d. That is,

$$\tilde{M}_Q = \begin{pmatrix} q_{00}^{d-1}q_{10} & q_{01}^{d-1}q_{11} \\ q_{00}q_{10}^{d-1} & q_{01}q_{11}^{d-1} \end{pmatrix}.$$

Given $\mathcal{T} \in V_{\text{diag}}$, $Q \bullet \mathcal{T} \in V_{\text{pmi}}$ if and only if $(t_0, t_d) \in \ker(M_Q)$. We are interested in $Q \in \mathrm{O}(2) \backslash \mathrm{SP}(2)$. If d is odd, M_Q has full rank for all $Q \in \mathrm{O}(2) \backslash \mathrm{SP}(2)$. If d is even, M_Q has full rank for all $Q \in \mathrm{O}(2) \backslash \mathrm{SP}(2)$ except if $q_{00} = \pm q_{01}$ and, in such a case, $\ker(M_Q) = \langle (1, -1) \rangle$. This concludes the proof for $n = 2$. We reduce the case $n \geq 3$ to the binary case by considering the normal form of orthogonal matrices and using the same reasoning as in the proof of Theorem 4.1. □

Remark 4.5. In [RSZ25] we also study tensors in $\mathbb{R}^{n_1 \times \cdots \times n_d}$ with an orthogonal basis of singular vector tuples. In that case, V_{pmi} is replaced by an analogous linear space V and the action of $\mathrm{O}(n)$ on $S^d(\mathbb{R}^n)$ is replaced by the action of $\mathrm{O}(n_1) \times \cdots \times \mathrm{O}(n_d)$ on $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_d}$. For $(\mathbb{R}^2)^{\otimes 3}$, the set of tensors with an orthogonal basis of singular vector tuples is precisely the set of odeco tensors, i.e., $V = V_{\text{diag}}$. Therefore, using the same reasoning as above, we can conclude that a tensor in $V \subset \mathbb{R}^{n_1 \times n_2 \times n_3}$ has a unique basis of orthogonal singular vector tuples if and only if at most one of its diagonal entries is zero. For $(\mathbb{R}^2)^{\otimes 4}$, we get that the tensors $\mathcal{T} \in V$ such that $(Q_1, \ldots, Q_4) \cdot \mathcal{T} \in V$ with some $Q_k \notin \mathrm{SP}(2)$ live in a four-dimensional variety consisting of 14 linear components:

$$\mathcal{T}_{0000} \pm \mathcal{T}_{0011} = \mathcal{T}_{0101} \mp \mathcal{T}_{0110} = \mathcal{T}_{1001} \mp \mathcal{T}_{1010} = \mathcal{T}_{1100} \pm \mathcal{T}_{1111} = 0,$$
$$\mathcal{T}_{0000} \pm \mathcal{T}_{0101} = \mathcal{T}_{0011} \mp \mathcal{T}_{0110} = \mathcal{T}_{1001} \mp \mathcal{T}_{1100} = \mathcal{T}_{1010} \pm \mathcal{T}_{1111} = 0,$$
$$\mathcal{T}_{0000} \pm \mathcal{T}_{0110} = \mathcal{T}_{0011} \mp \mathcal{T}_{0101} = \mathcal{T}_{1010} \mp \mathcal{T}_{1100} = \mathcal{T}_{1001} \pm \mathcal{T}_{1111} = 0,$$
$$\mathcal{T}_{0000} \pm \mathcal{T}_{1001} = \mathcal{T}_{0011} \mp \mathcal{T}_{1010} = \mathcal{T}_{0101} \mp \mathcal{T}_{1100} = \mathcal{T}_{0110} \pm \mathcal{T}_{1111} = 0,$$
$$\mathcal{T}_{0000} \pm \mathcal{T}_{1010} = \mathcal{T}_{0011} \mp \mathcal{T}_{1001} = \mathcal{T}_{0110} \mp \mathcal{T}_{1100} = \mathcal{T}_{0101} \pm \mathcal{T}_{1111} = 0,$$
$$\mathcal{T}_{0000} \pm \mathcal{T}_{1100} = \mathcal{T}_{0101} \mp \mathcal{T}_{1001} = \mathcal{T}_{0110} \mp \mathcal{T}_{1010} = \mathcal{T}_{0011} \pm \mathcal{T}_{1111} = 0,$$
$$\mathcal{T}_{0000} \pm \mathcal{T}_{1111} = \mathcal{T}_{0011} \pm \mathcal{T}_{1100} = \mathcal{T}_{0101} \pm \mathcal{T}_{1010} = \mathcal{T}_{0110} \pm \mathcal{T}_{1001} = 0.$$

5. Other relaxations of independence

Here, we prove Theorem 1.4; i.e., we show generic identifiability of component analysis for any condition on the cumulants of the source variables that is weaker than independence (V_{diag}) and stronger than pairwise mean independence (V_{pmi}). Then we study examples of

such distributions. We close the section by proving Theorem 1.5; i.e., PMICA becomes generically unidentifiable if one mean independence assumption is dropped.

PROOF OF THEOREM 1.4. We use the same idea as in the proof of Theorem 1.3. The d-th cumulant tensor of a sufficiently general independent distribution is generic as a tensor in V_{pmi} in the sense of Theorem 3.2, by Lemma 3.3. We are assuming that $V_{\mathrm{diag}} \subseteq V$. Hence, if $\kappa_d(\mathbf{s})$ is sufficiently general in V, it is also sufficiently general in V_{pmi}. Therefore, the result follows from Theorem 1.3.   □

Remark 5.1. In Theorem 1.4 we assume that $\mathrm{Cov}(\mathbf{s}) = I$ so that we can restrict our search space to $\mathrm{O}(n)$ by applying the whitening transformation to $\mathbf{x}$. This assumption was not needed in Theorem 1.3 because PMI implies uncorrelatedness and scaling $\mathbf{s}$ does not change membership in the model, since V_{pmi} is given by zero restrictions. Similarly, this assumption is not needed in Theorem 1.4 whenever $V \subset S^d(\mathbb{R}^n)$ is given by zero restrictions.

Lemma 5.2. *For $d \geq 3$, any real number can be obtained as the d-th cumulant of some probability distribution.*

PROOF. Given a random variable z and $\alpha \in \mathbb{R}$ we have $\kappa_d(\alpha z) = \alpha^d \kappa_d(z)$, so it suffices to study the sign of cumulants. Let z be a Bernoulli random variable with parameter p, and denote its d-th cumulant as $\kappa_d(p)$. Then $\kappa_d(p)$ is a polynomial of degree d in p that satisfies the following recursion: $\kappa_1 = p$ and $\kappa_{d+1} = p(1-p)\frac{d}{dp}\kappa_d$. Fix $d \geq 2$, then $\kappa_d(0) = \kappa_d(1) = 0$ and $\kappa_d(p)$ is not identically zero, so there exist $p^+, p^- \in (0,1)$ such that $\frac{d}{dp}\kappa_d(p^+) > 0$ and $\frac{d}{dp}\kappa_d(p^-) < 0$, by the mean value theorem. Therefore, $\kappa_{d+1}(p^+) > 0$ and $\kappa_{d+1}(p^-) < 0$.   □

Proposition 5.3. *Any symmetric tensor $\mathcal{T} \in S^d(\mathbb{R}^n)$ of order $d \geq 3$ can be obtained as a cumulant tensor of some probability distribution.*

PROOF. There is no sign condition on univariate cumulants for $d \geq 3$ (only for $d = 2$ where the variance must be non-negative), by Lemma 5.2. Given any $\mathcal{T} \in S^d(\mathbb{R}^n)$, consider a symmetric decomposition $\mathcal{T} = \sum_{i=1}^r \lambda_i \mathbf{v}_i^{\otimes d}$, which always exists by [CGLM08, Lemma 4.2]. Construct r independent random variables $z_1, \ldots, z_r$ such that $\kappa_d(z_i) = \lambda_i$ and define the random vector $\mathbf{y} = \sum_{i=1}^r z_i \mathbf{v}_i$. Then $\kappa_d(\mathbf{y}) = \sum_{i=1}^r \kappa_d(z_i \mathbf{v}_i) = \mathcal{T}$.   □

The previous proposition implies that it is not necessary to assume that $V_{\mathrm{diag}} \subseteq V$ in Theorem 1.4, and the statement holds as long as V contains sufficiently generic tensors in V_{pmi}. For example, one can take V to be the subset of diagonal tensors whose first diagonal entry is zero, which contains sufficiently general tensors in V_{pmi} for $d = 3, 4$. This recovers the classical ICA identifiability condition that one source may be Gaussian [Com94].

When we talk about a tensor $\mathcal{T}$ being generic in $V \subset S^d(\mathbb{R}^n)$ in the context of component analysis, we mean that $\{Q \in \mathrm{O}(n) \mid Q \bullet \mathcal{T} \in V\} \subseteq \mathrm{SP}(n)$, following [MZ24]. One benefit of Theorem 1.4 is that one does not need to study the genericity conditions for each $V \subseteq V_{\mathrm{pmi}}$ as long as $V_{\mathrm{diag}} \subseteq V$. Hence, our identifiability result can be applied to any domain-specific family of distributions containing generic independent distributions (in the sense of Theorem 4.4) and having cumulant tensors in $V_{\mathrm{pmi}} \subset S^d(\mathbb{R}^n)$ for some $d \geq 3$. For example, this can be applied to a parametric family of distributions satisfying such properties, which can be useful in applications. We formalize this discussion as follows.

Corollary 5.4. *Let $\mathcal{F}$ be a family of probability distributions that contains generic independent distributions and whose cumulants tensors lie in $V_{\mathrm{pmi}} \subset S^d(\mathbb{R}^n)$ for some $d \geq 3$.*

Consider the model $\mathbf{x} = A\mathbf{s}$ *where* $A \in \mathbb{R}^{n \times n}$ *is invertible and* $\mathbf{s} \in \mathcal{F}$ *is sufficiently general.* *Then* A *is identifiable from* $\kappa_d(\mathbf{x})$ *(up to permutation and scaling of columns).*

The following is a consequence of Theorem 1.4; see [RSZ25, Proposition 7.1] for details.

Corollary 5.5. *Let* $V_{\mathrm{diag}} \subseteq V \subseteq V_{\mathrm{pmi}}$. *Then,*

$$\dim(\mathrm{O}(n) \bullet V) = \dim(V) + \dim(\mathrm{O}(n)) = \dim(V) + \binom{n}{2}.$$

When we recover the mixing matrix from cumulants, the ICA model corresponds to symmetric odeco tensors $\mathrm{O}(n) \bullet V_{\mathrm{diag}}$ and the PMICA model corresponds to tensors with and orthogonal basis of eigenvectors $\mathrm{O}(n) \bullet V_{\mathrm{diag}}$. Given $d \geq 3$, $\dim(\mathrm{O}(n) \bullet V_{\mathrm{diag}}) = \binom{n+1}{2} \asymp n^2$ and $\dim(\mathrm{O}(n) \bullet V_{\mathrm{pmi}}) = \binom{n+d-1}{d} - \binom{n}{2} \asymp n^d$. This shows how PMICA is more expressive than ICA: given a cumulant tensor $\kappa_d(\mathbf{x})$ in practice, we expect to approximate it better with $\mathrm{O}(n) \bullet V_{\mathrm{pmi}}$ than with $\mathrm{O}(n) \bullet V_{\mathrm{diag}}$.

5.1. Examples between independence and PMI. We give examples of conditions that are stronger than PMI but weaker than independence, so Theorem 1.4 applies.

Pairwise mean independence is stronger than uncorrelatedness: if z_1 is mean independent of z_2, then $\mathbb{E}(z_1 z_2) = \mathbb{E}(\mathbb{E}(z_1 \mid z_2) z_2) = \mathbb{E}(\mathbb{E}(z_1) z_2) = \mathbb{E}(z_1)\mathbb{E}(z_2)$. To see that both notions are not equivalent, take the vector (z, z^2) where z is standard Gaussian. In this case $\mathrm{cov}(z, z^2) = \mathbb{E}(z^3) = 0$ but $\mathbb{E}(z^2 \mid z) = z^2$, which is not equal to $\mathbb{E}(z^2) = 1$ almost surely.

5.1.1. *Reflectional symmetries on cumulants.* The conditions in [MZ24] (general common covariance, multiple scaled elliptical distribution, mean independent) are special cases of our setting, since the set of reflectionally invariant tensors V_{refl} is included in V_{pmi}. Indeed, $V_{\mathrm{refl}} = \{\mathcal{T} \in S^d(\mathbb{R}^n) \mid \mathcal{T} = D \bullet \mathcal{T} \text{ for all } D \in \mathrm{diag}(\{\pm 1\}^n)\}$, where $\mathrm{diag}(\{\pm 1\}^n)$ is the set of diagonal matrices with ± 1 on the diagonal [MZ24, Section 5], so in $S^d(\mathbb{R}^n)$ we get

$$\dim(V_{\mathrm{refl}}) = \begin{cases} 0 & \text{if } d = 2k+1 \\ \binom{n+k-1}{k} & \text{if } d = 2k, \end{cases}$$

which is smaller than $\dim(V_{\mathrm{pmi}}) = \binom{n+d-1}{d} - n(n-1)$.

5.1.2. *Mean independence.* Define

$$V_{\mathrm{mi}} := \{\mathcal{T} \in S^d(\mathbb{R}^n) \mid \mathcal{T}_{i,j_2,\ldots,j_d} = 0 \text{ if } i \neq j_2, \ldots, j_d\}.$$

In words, $\mathcal{T}_{j_1,j_2,\ldots,j_d} = 0$ if there is an index that appears only once in the tuple $(j_1, j_2, \ldots, j_d)$. Theorem 2.3 holds replacing PMI by mean independence and V_{pmi} by V_{mi}. To see this, we can follow the proof of Theorem 2.3 replacing (3) with

$$(6) \quad \mathbb{E}(z_i z_{j_2} \cdots z_{j_d}) = \mathbb{E}\big(z_{j_2} \cdots z_{j_d} \mathbb{E}(z_i \mid z_{\backslash i})\big) = \mathbb{E}(z_i)\mathbb{E}(z_{j_2} \cdots z_{j_d}) \quad (i \neq j_2, \ldots, j_d).$$

In (4) specialize $i_1, i_2, \ldots, i_d$ to $i, j_2, \ldots, j_d$ with $i \neq j_2, \ldots, j_d$ with the same special partition π_0 to conclude that with mean independence $\kappa_d(\mathbf{x}) \in V_{\mathrm{pmi}}^{d,n}$.

We have the containment $V_{\mathrm{mi}}^d \subseteq V_{\mathrm{pmi}}^d$ with strict inclusion for $d \geq 3$ provided $n \geq 3$, so mean independence implies PMI. This property can also be seen as follows.

Proposition 5.6. *If* $\mathbf{z} = (z_1, \ldots, z_n)$ *is mean independent, it is pairwise mean independent.*

PROOF. By the tower property of conditional expectation, for any $i \neq j$, $\mathbb{E}(z_i \mid z_j) = \mathbb{E}(\mathbb{E}(z_i \mid z_{\backslash i}) \mid z_j) = \mathbb{E}(z_i)$ almost surely. $\quad \square$

5.1.3. *Pairwise independence.* A random vector $\mathbf{z}$ is *pairwise independent* if $z_i \perp\!\!\!\perp z_j$ for all $i \neq j$. This is equivalent to $\kappa(\mathbf{z})_{i_1,\ldots,i_d} = 0$ whenever $|\{i_1,\ldots,i_d\}| = 2$, provided that $K_{\mathbf{z}}(\mathbf{t})$ is sufficiently smooth around zero. Let

$$V_{2-\text{indep}}^d = \{\mathcal{T} \in S^d(\mathbb{R}^n) \mid \mathcal{T}_{i_1,\ldots,i_d} = 0 \text{ if } |\{i_1,\ldots,i_d\}| = 2\}.$$

Let $\mathcal{I}_{2-\text{indep}} = \{\mathbf{i} \in [n]^d \mid 1 \leq i_1 \leq \cdots \leq i_d \leq n, |\{i_1,\ldots,i_d\}| = 2\}$. Then $\text{codim}(V_{2-\text{indep}}^d) = |\mathcal{I}_{2-\text{indep}}| = (d-1)\binom{n}{2}$ and $V_{2-\text{indep}}^d \subseteq V_{\text{PMI}}^d$ for all $d \geq 2$.

5.1.4. *k-wise independence.* . Let $k \in [n]$. A random vector $\mathbf{z}$ is *k-wise independent* if every set of k distinct entries of $\mathbf{z}$ are jointly independent.

Remark 5.7. Independence $\implies$ k-wise independence $(k \geq 2)$ $\implies$ pairwise independence. More precisely, k-wise independence $\implies$ $(k-1)$-independence.

If $K_{\mathbf{z}}(\mathbf{t})$ is sufficiently smooth around zero, then k-wise independence is equivalent to $\kappa(\mathbf{z})_{i_1,\ldots,i_d} = 0$ whenever $2 \leq |\{i_1,\ldots,i_d\}| \leq k$. Let

$$V_{k-\text{indep}}^d = \{\mathcal{T} \in S^d(\mathbb{R}^n) \mid \mathcal{T}_{i_1,\ldots,i_d} = 0 \text{ if } 2 \leq |\{i_1,\ldots,i_d\}| \leq k\}$$

Let $\mathcal{I}_{k-\text{indep}} = \{\mathbf{i} \in [n]^d \mid 1 \leq i_1 \leq \cdots \leq i_d \leq n, 2 \leq |\{i_1,\ldots,i_d\}| \leq k\}$. Then $\text{codim}(V_{k-\text{indep}}^d) = |\mathcal{I}_{k-\text{indep}}| = \sum_{j=2}^{k} \binom{d-1}{j-1}\binom{n}{j}$ and $V_{k-\text{indep}}^d \subseteq V_{\text{PMI}}^d$ for all $d \geq 2$.

Remark 5.8. We have $V_{k-\text{indep}}^d = V_{(k+1)-\text{indep}}^d \subset S^d(\mathbb{R}^n)$ whenever $k \geq \min\{d,n\}$. That is, the d-th order cumulant does not distinguish between k-wise independence and independence whenever $k \geq \min\{d,n\}$.

5.1.5. *Correlation of energies.* Let $z_i = \sigma_i \varepsilon_i$, where σ_i, ε_i are random variables, $\varepsilon_i \perp\!\!\!\perp \varepsilon_j$ for all $i \neq j$ and $\sigma_i \perp\!\!\!\perp \varepsilon_j$ for all i, j. That is, $z_i \perp\!\!\!\perp z_j \mid \sigma_i, \sigma_j$ but $z_i \not\perp\!\!\!\perp z_j$. Assume that $\sigma_i > 0$ a.s, $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\varepsilon_i^2) = 1$. Hence, $\mathbb{E}(z_i) = 0$. Topographic ICA (tICA) [HHI01] is a particular case of this model.

Proposition 5.9. *Correlation of energies implies pairwise mean independence.*

PROOF. $\mathbb{E}(z_i \mid z_j) = \mathbb{E}(\sigma_i \varepsilon_i \mid \sigma_j \varepsilon_j) = \mathbb{E}(\sigma_i \mid \sigma_j \varepsilon_j)\mathbb{E}(\varepsilon_i) = 0 = \mathbb{E}(z_i)$. $\square$

Example 5.10 (Broadcasting on trees). Let $T = (V, E)$ be a rooted tree with root denoted by 0 and non-root leaves denoted by $\{1,\ldots,n\}$. For a vertex $v \in V$, we write $u \preceq v$ if u lies on the unique path from the root 0 to v. Let $\{\tau_v \mid v \in V\}$ be independent random variables with zero mean. Consider the broadcasting process $\{z_v : v \in V\}$ defined as: $z_v = \prod_{u \preceq v} \tau_u$. Then $\mathbf{z} = (z_1,\ldots,z_n)$ is pairwise mean independent.

5.1.6. *Spherical distributions.* In spherical distributions the components are mean independent and pairwise mean independent but not independent unless Z is Gaussian; see [Kel70, RZ21]. However, cumulants of spherical distributions are insufficiently general in V_{pmi}. This is because their distribution is preserved under rotation.

5.2. Generic non-identifiability beyond pairwise mean independence. We prove Theorem 1.5. Being mean independent is not symmetric: $\mathbb{E}(z_1 \mid z_2) = \mathbb{E}(z_1)$ does not imply that $\mathbb{E}(z_2 \mid z_1) = E(z_2)$. For example, take $z_1 = U(-1,1)$ and $z_2 = z_1^2$. Then $\mathbb{E}(z_1 \mid z_2) = 0 = \mathbb{E}(z_1)$ but $\mathbb{E}(z_2 \mid z_1) = z_1^2 \neq 1/3 = \mathbb{E}(z_2)$ almost surely. If z_2 is not mean independent of z_1 we cannot guarantee that $\kappa(\mathbf{x})_{21\ldots1}$ is zero. The following shows that dropping one mean independence condition leads to generic unidentifiability of the component analysis model. That is, if s_1 is mean independent of s_2 but s_2 is not mean independent of s_1, then there are many rotations that preserve this property.

Lemma 5.11. *Let $d \geq 3$ and consider a linear space $V_{\mathrm{pmi}} \subsetneq V \subseteq S^d(\mathbb{R}^n)$ given by zero restrictions. Let $\mathcal{T} \in V$ be generic. Then, there exists $Q \in \mathrm{O}(n) \backslash \mathrm{SP}(n)$ such that $Q \bullet \mathcal{T} \in V$.*

PROOF. Consider first $n = 2$. Then $V_{\mathrm{pmi}} = \{\mathcal{T} \in S^d(\mathbb{R}^2) \mid \mathcal{T}_{12\ldots2} = \mathcal{T}_{21\ldots1} = 0\}$. Suppose that $V = \{\mathcal{T} \in S^d(\mathbb{R}^2) \mid \mathcal{T}_{12\ldots2} = 0\}$, the argument for the other cases is analogous. Then V is the set of symmetric tensors with e_2 as an eigenvector, so $\mathrm{O}(2) \bullet V = S^d(\mathbb{R}^2)$ because every symmetric tensor has an eigenvector. A generic tensor in $S^d(\mathbb{R}^2)$ has $\frac{(d-1)^2-1}{d-1}$ distinct complex eigenvectors [CS13, Theorem 5.5]. At least two of these eigenvectors are real, corresponding to the maximizer and minimizer of $\mathcal{T}(x, \ldots, x)$ subject to $\|x\| = 1$. Moreover, these two real eigenvectors are not orthogonal to each other by genericity of $\mathcal{T}$ and [RSZ25, Proposition 7.1.]. This means that a generic fiber of the parametrization $\phi : \mathrm{O}(2) \times V \to S^d(\mathbb{R}^n), (Q, \mathcal{T}) \mapsto Q \bullet \mathcal{T}$ is finite, but it is not included in $\mathrm{SP}(2) \times V$. Hence, for a generic $\mathcal{T} \in V$, there exists $Q \in \mathrm{O}(2) \setminus \mathrm{SP}(2)$ such that $Q \bullet \mathcal{T} \in V$.

Next, let $n \geq 3$ and let $\mathcal{I}_{\mathrm{pmi}} = \{(i, j, \ldots, j) \in [n^d] \mid i \neq j\}$. Then $V_{\mathrm{pmi}} = \{\mathcal{T} \in S^d(\mathbb{R}^n) \mid \mathcal{T}_{\mathbf{i}} = 0$ for all $\mathbf{i} \in \mathcal{I}_{\mathrm{pmi}}\}$. Let $\mathcal{I} = \mathcal{I}_{\mathrm{pmi}} \setminus \{(2, 1, \ldots, 1)\}$ and let $V = \{\mathcal{T} \in S^d(\mathbb{R}^n) \mid \mathcal{T}_{\mathbf{i}} = 0$ for all $\mathbf{i} \in \mathcal{I}\}$. By the binary case, given a generic $\mathcal{T} \in V$, there exists a matrix

$$
Q = \begin{pmatrix} \tilde{Q} & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}
$$

with $\tilde{Q} \in \mathrm{O}(2) \setminus \mathrm{SP}(2)$ such that $Q \bullet \mathcal{T} \in V$. $\qquad\square$

PROOF OF THEOREM 1.5. Without loss of generality, suppose s_2 is not mean independent of s_1. Here $\mathbf{s}$ being general means that for all $d \geq 3$, $\kappa_d(\mathbf{s})$ is generic in $V = \{\mathcal{T} \in S^d(\mathbb{R}^n) \mid \mathcal{T}_{ij\ldots j} = 0$ for all $i \neq j, (i, j) \neq (2, 1)\}$, so the statement follows by Lemma 5.11. $\qquad\square$

Theorem 1.5 says that one cannot identify A from $\kappa_d(\mathbf{x})$ if if one does not imposes all the zero-restrictions $\kappa_d(\mathbf{s})$ coming from mean independence. Generic in this context means that $\kappa_d(\mathbf{s}) \notin V_{\mathrm{pmi}}$ for any $d \geq 3$. However, Theorem 1.4 says that it is enough to have a generic $\kappa_d(\mathbf{s}) \in V_{\mathrm{pmi}}$ for only one $d \geq 3$ to have identifiability of the model, even if $\mathbf{s}$ is not PMI. The following is an example of such a distribution.

Example 5.12. Let $\mathbf{s} = (s_1, s_2)$ with $s_1 \sim \mathcal{N}(0, 1)$ and $s_2 = (s_1^2 - 1)/\sqrt{2}$. We have $\mathbb{E}(s_1) = \mathbb{E}(s_2) = 0$ and $\mathbb{E}(s_1 \mid s_2) = 0$ but $\mathbb{E}(s_2 \mid s_1) = (s_1^2 - 1)/\sqrt{2} \neq 0$ almost surely, so $\mathbf{s}$ is not PMI. However, $\kappa_4(\mathbf{s})_{1222} = \kappa_4(\mathbf{s})_{2111} = 0$, so $\kappa_4(\mathbf{s}) \in V_{\mathrm{pmi}}$. Moreover, $\kappa_4(\mathbf{s})$ is sufficiently general in V_{pmi} in the context of Theorem 1.4 because $\kappa_4(\mathbf{s})_{1111} = 0 \neq \kappa_4(\mathbf{s})_{2222} = 12$; see Theorem 4.1. Therefore, in the model $\mathbf{x} = A\mathbf{s}$, A can be identified from $\kappa_4(\mathbf{x})$, up to permutation and sign flip of columns.

6. A minimum-distance estimator: consistency and finite-sample behavior

Our identifiability analysis applies to any model obtained by restricting the cumulant at some fixed $d \geq 3$ to lie in a linear subspace $V \subset S^d(\mathbb{R}^n)$ with

$$
(7) \qquad\qquad V_{\mathrm{diag}}^{d,n} \subseteq V \subseteq V_{\mathrm{pmi}}^{d,n}.
$$

Our two main examples are ICA and PMICA. We observe $\mathbf{x} = A\mathbf{s}$, where $\mathbf{s}$ is centered with $\mathrm{Cov}(\mathbf{s}) = I$, and we assume that $\kappa_d(\mathbf{s})$ lies in V and is generic in the sense of Section 4.

6.1. Minimum distance estimator. Consider i.i.d. samples $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^n$ from $\mathbf{x}$. Stack the samples to form the matrix $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times n}$ and whiten them to obtain X_w. Then $\mathbf{x}_w = \tilde{A}\mathbf{s}$ with $\tilde{A} \in \mathrm{O}(n)$, as explained in Section 3.

For $d \geq 3$, let $\widehat{h}_d(\mathbf{x}_w)$ denote an order-d estimator, for example the sample moment tensor $\widehat{\mu}_d(\mathbf{x}_w)$, sample cumulant tensor $\widehat{\kappa}_d(\mathbf{x}_w)$, or an alternative cumulant estimator such as the order-d k-statistics; see [McC18, Chapter 4]. Let $h_d(\mathbf{x}_w)$ be the population version.

Population formulation. Let Π_W be the orthogonal projector (with the Frobenius inner product) onto a subspace $W \subset S^d(\mathbb{R}^n)$. For a fixed V satisfying (7) define

$$g(Q) := \Pi_{V^\perp}\big(Q^\top \bullet h_d(\mathbf{x}_w)\big), \qquad Q \in \mathrm{O}(n).$$

Using multilinearity of moments and cumulants (Lemma 3.1),

$$g(Q) = \Pi_{V^\perp}\big((Q^\top \tilde{A}) \bullet h_d(\mathbf{s})\big).$$

The squared Euclidean distance between the tensor $Q^\top \bullet h_d(\mathbf{x}_w)$ and the linear space V is found by the optimization problem

$$(8) \qquad \text{minimize} \quad F(Q) := \|g(Q)\|_{\mathrm{F}}^2, \qquad Q \in \mathrm{O}(n).$$

The following lemma follows from Theorem 1.4.

Lemma 6.1 (Population identification). *Suppose $h_d(\mathbf{s})$ is generic in V. Then $g(Q) = 0$ if and only if $Q^\top \tilde{A} \in \mathrm{SP}(n)$. Hence, the minimizers of F are $\tilde{A}P$ with $P \in \mathrm{SP}(n)$.*

Letting $\Sigma = \mathrm{Cov}(\mathbf{x})$, recall from Section 3 that $\tilde{A} = \Sigma^{-1/2}A$ and so we recover A up to signed permutation via

$$(9) \qquad\qquad \Sigma^{1/2}\tilde{A}P = AP, \qquad P \in \mathrm{SP}(n).$$

Remark 6.2 (Equivalent viewpoints). Lemma 6.1 shows that the minimizers Q of F are signed permutations of $\tilde{A}$. Equivalently, the columns of $Q^\top \tilde{A}$ are an orthonormal set of eigenvectors of the tensor $h_d(\mathbf{s})$. From a variational perspective, the eigenvectors are the stationary points of

$$u \longmapsto \langle h_d(\mathbf{s}), u^{\otimes d}\rangle$$

on the unit sphere. See [RSZ25, Sec. 4] for a discussion of this eigenvector formulation.

Sample formulation. Mimicking the population construction, we estimate $\tilde{A}$ by solving

$$(10) \qquad \widehat{Q} \in \underset{Q \in \mathrm{O}(n)}{\arg\min} \; F_N(Q) := \|g_N(Q)\|_{\mathrm{F}}^2, \qquad g_N(Q) := \Pi_{V^\perp}\big(Q^\top \bullet \widehat{h}_d(\mathbf{x}_w)\big).$$

We solve this optimization problem using Riemannian gradient descent (RGD) on $\mathrm{O}(n)$. Analogously to (9), to estimate A, compute the sample covariance $\widehat{\Sigma}$ of the data and set

$$\widehat{A}_N = \widehat{\Sigma}^{1/2}\widehat{Q}.$$

6.2. Large-sample theory. The criterion $F_N(Q) = \|\Pi_{V^\perp}(Q^\top \bullet \widehat{h}_d)\|_{\mathrm{F}}^2$ leads to the Generalized Method of Moments (GMM) estimator, with moment conditions the defining equations of V. The standard GMM results apply; we state them for completeness and refer to [MZ24, Propositions 6.2–6.3] and [Han82] for proofs, generalizations, and discussion.

Proposition 6.3 (Consistency). *Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be i.i.d. samples from $\mathbf{x} = A\mathbf{s}$. Assume that $\kappa_2(\mathbf{s}) = I_n$, $\kappa_d(\mathbf{s})$ is generic in V, and $\mathbb{E}\|\mathbf{x}\|^d < \infty$. Then $\widehat{A}_N \to_p AP$ for some $P \in \mathrm{SP}(n)$.*

Let $\mathrm{vec}g : \mathbb{R}^{n^2} \to \mathbb{R}^{\dim(V^\perp)}$ denote the map g after vectorizing its domain and codomain. Define $\mathrm{vec}g_N$ similarly. Let $G(A) \in \mathbb{R}^{n^2 \times \dim(V^\perp)}$ denote the Jacobian of $\mathrm{vec}g$ at A and define $\Xi = \lim_{N \to \infty} \mathrm{Var}(\sqrt{N} \, \mathrm{vec}g_N(AP))$.

Proposition 6.4 (Asymptotic normality)**.** *Under the assumptions of Proposition 6.3, assume additionally that $\mathbb{E}\|\mathbf{x}\|^{2d} < \infty$. Write $G := G(AP)$. Then*

$$\sqrt{N} \, \mathrm{vec}(\widehat{A}_N - AP) \ \to \mathcal{N}\Big(0, \ (G^\top G)^{-1} G^\top \, \Xi \, G \, (G^\top G)^{-1}\Big).$$

This shows that our estimator has the standard good asymptotic properties: consistency and asymptotic normality. Next, we provide finite sample bounds for sub-Gaussian data.

6.3. Finite-sample error bounds under sub-Gaussianity. We develop basic finite sample analysis. We focus on moments to keep things simple; c.f. Corollary 2.4. The proof of Proposition 6.7 and technical material used in this section is in Section A.

Assumption 6.5 (Sub-Gaussian data)**.** The data $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are whitened, so $\mathbf{x} \in \mathbb{R}^n$ satisfies $\mathbb{E}[\mathbf{x}] = 0$ and $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = I$, and $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are i.i.d. copies of $\mathbf{x}$. We also assume $\mathbf{x}$ is a sub-Gaussian vector, that is, there exists $\sigma > 0$ such that $\langle u, \mathbf{x} \rangle$ is σ-sub-Gaussian for every unit vector $u \in \mathbb{R}^n$.

Define the d-th order sample moment tensor $\widehat{\mu}_d = \frac{1}{N} \sum_{r=1}^{N} \mathbf{x}_r^{\otimes d}$ and the population moment tensor $\mu_d = \mathbb{E}[\mathbf{x}^{\otimes d}]$. We bound the spectral norm (see [Lim05])

$$(11) \qquad \|\widehat{\mu}_d - \mu_d\|_2 := \max_{\|u\|=1} \langle \widehat{\mu}_d - \mu_d, u^{\otimes d} \rangle.$$

For any unit vector $u \in \mathbb{R}^n$, define $\mathbf{z} = u^\top \mathbf{x}$ and $\mathbf{z}_i = u^\top \mathbf{x}_i$ for $i = 1, \ldots, N$. Then

$$\langle \widehat{\mu}_d - \mu_d, u^{\otimes d} \rangle \ = \ \frac{1}{N} \sum_{i=1}^{N} \left(z_i^d - \mathbb{E}[\mathbf{z}^d] \right) \ = \ \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{z}_i^d - \mathbb{E}[\mathbf{z}_i^d] \right).$$

This allows the use of concentration of measure techniques to bound $\|\widehat{\mu}_d - \mu_d\|_2$. The next result follows directly from [AGCSA25, Theorem 2.1] (with $H = \mathbb{R}^n$ and $\Sigma = I$).

Proposition 6.6 (Concentration bound)**.** *Under Assumption 6.5, there is a constant $C > 0$ (depending only on d and the sub-Gaussian parameter σ) such that for all $\tau > 0$,*

$$\mathrm{Pr}\left(\|\widehat{\mu}_d - \mu_d\|_2 \ \leq \ C \left(\sqrt{\frac{n}{N}} \ + \ \frac{n^{d/2}}{N} + \sqrt{\frac{\tau}{N}} + \frac{\tau^{d/2}}{N} \right) \right) \ \geq \ 1 - e^{-\tau}.$$

The bounds on $\|\widehat{\mu}_d - \mu_d\|_2$ imply bounds on the rotation recovery $\|\widehat{Q} - Q\|_\mathrm{F}$ in the objectives of our estimation problem in (8) and (10). That is, since $\widehat{\mu}_d$ is close to μ_d, the optima of F_N should be close to those of F. Quantifying this relies on the Hessian of F.

Extend F to be a function on $\mathbb{R}^{n \times n}$. Its derivative $\mathrm{D}F(Q)$ at $Q \in \mathrm{O}(n)$ is a linear functional on $\mathbb{R}^{n \times n}$. Similarly, the second derivative $\mathrm{D}^2 F(Q)$ is a linear map $\mathrm{D}^2 F(Q) : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$. Restricting $\mathrm{D}^2 F(Q)[\Delta]$ to Δ in the tangent space $U = T_Q \mathrm{O}(n)$ and projecting the image to U, gives the Riemannian Hessian $\mathrm{Hess}F(Q) : U \to U$. In other words,

$$\mathrm{Hess}F(Q)[\Delta] \ := \ \Pi_U(\mathrm{D}^2 F(Q)[\Delta]) \qquad \text{for } \Delta \in U.$$

Proposition 6.7 (Parameter error in Frobenius norm). *Let $\widehat{Q} \in \mathrm{O}(n)$ be any empirical minimizer of F_N in (10). Assume there is a population minimizer $Q^\star$ of $F(Q) = \|g(Q)\|_{\mathrm{F}}^2$ such that the following curvature condition holds at every point of its orbit $Q^\star \mathrm{SP}(n)$: there exists $\kappa > 0$ with*

$$\langle \mathrm{Hess}F(Q)[\Delta], \ \Delta \rangle \ \geq \ \kappa \, \|\Delta\|_{\mathrm{F}}^2 \quad \text{for all } \Delta \in T_Q\mathrm{O}(n).$$

Then, there exists $R \in \mathrm{SP}(n)$ such that

$$(12) \qquad \|\widehat{Q} - Q^\star R\|_{\mathrm{F}} \ \leq \ \frac{C_d}{\kappa} \left(d \, \|\mu_d\|_{\mathrm{F}} \ + \ d \, \|\widehat{\mu}_d - \mu_d\|_{\mathrm{F}} \right) \|\widehat{\mu}_d - \mu_d\|_{\mathrm{F}},$$

where $C_d > 0$ depends only on d. In particular, with probability at least $1 - e^{-\tau}$,

$$\|\widehat{Q} - Q^\star R\|_{\mathrm{F}} \ \lesssim_d \ \frac{n^{\frac{d-1}{2}}}{\kappa} \, \|\mu_d\|_{\mathrm{F}} \left(\sqrt{\frac{n}{N}} \ + \ \frac{n^{d/2}}{N} \ + \ \sqrt{\frac{\tau}{N}} \ + \ \frac{\tau^{d/2}}{N} \right),$$

where the implicit constant depends only on d and on the sub-Gaussian parameter of $\mathbf{x}$.

7. Local versus global optima under pairwise mean independence

We now focus on a misspecified scenario: we assume $V = V_{\mathrm{diag}}^{d,n}$, but the true tensor lies in the larger space, $V_{\mathrm{pmi}}^{d,n}$. When $V = V_{\mathrm{diag}}$, minimizing $\Pi_{V^\perp}\big(Q^\top \bullet h_d(\mathbf{s})\big)$ over $Q \in \mathrm{O}(n)$ as in (8) is equivalent to

$$(13) \qquad\qquad \text{maximize} \ \sum_{i=1}^n \big\langle h_d(\mathbf{s}), \ q_i^{\otimes d} \big\rangle^2, \qquad Q \in \mathrm{O}(n),$$

where $q_1, \ldots, q_n$ are the columns of Q. In the true ICA setting $h_d(\mathbf{s}) \in V_{\mathrm{diag}}$, this is maximized at matrices Q such that $Q^\top \tilde{A} \in \mathrm{SP}(n)$, by Lemma 6.1. In other words, the procedure recovers the correct mixing matrix $\tilde{A}$ up to signed permutation.

To simplify notation we work at the population level and, without loss of generality, set $\tilde{A} = I$; i.e., $\mathbf{x} = \mathbf{s}$. With this convention, the identity I (and any signed permutation $P \in \mathrm{SP}(n)$) is a ground-truth rotation: sources are exactly recovered when the rows of Q are the basis vectors e_i up to sign. We decompose

$$h_d(\mathbf{s}) = \kappa = \kappa_{\mathrm{diag}} + \kappa_{\mathrm{off}}, \qquad \kappa_{\mathrm{diag}} = \sum_{a=1}^n \lambda_a e_a^{\otimes d}, \quad \lambda_a = \kappa_d(s_a),$$

where κ_{off} collects the PMI off-diagonal entries. That is, $\kappa_{\mathrm{diag}} = \Pi_{V_{\mathrm{diag}}}(\kappa)$ and $\kappa_{\mathrm{off}} = \Pi_{V_{\mathrm{diag}}^\perp}(\kappa)$ For $Q \in \mathrm{O}(n)$ we write

$$\Phi_{\mathrm{diag}}(Q) := \sum_{i=1}^n \big\langle \kappa_{\mathrm{diag}}, \ q_i^{\otimes d} \big\rangle^2,$$

so that the objective in (13) equals $\Phi_{\mathrm{diag}}(Q)$ when $\kappa_{\mathrm{off}} = 0$.

We show that I is a stationary point of (13). We write $Q = I + H$ with H an infinitesimal perturbation in the tangent space to $\mathrm{O}(n)$ at I ($H + H^\top = 0$, and hence H has zeros on the diagonal). Then

$$((I + H) \bullet \kappa)_{i\cdots i} = \kappa_{i\cdots i} + \sum_{\ell \neq i} H_{i\ell}\kappa_{\ell i\cdots i} + \cdots + \sum_{\ell \neq i} H_{i\ell}\kappa_{i\cdots i\ell} + o(\|H\|_{\mathrm{F}}).$$

If $\kappa \in V_{\mathrm{pmi}}$, then all terms with exactly one index $\ell \neq i$ vanish, so

$$((I + H) \bullet \kappa)^2_{i \cdots i} - (I \bullet \kappa)^2_{i \cdots i} = o(\|H\|_{\mathrm{F}}).$$

Hence I (and every $P \in \mathrm{SP}(n)$) is a stationary point of $F(Q)$ under PMI.

Our simulations in Section 8 show that this first order analysis can be misleading: as soon as $\|\kappa_{\mathrm{off}}\|_{\mathrm{F}}$ is moderately large, the global maximizer of F may move away from $\mathrm{SP}(n)$, and the landscape can develop spurious local maxima and saddle points.

We now consider local versus global behavior for two random variables. To make our first order analysis more explicit, consider $n = 2$ and $Q = R(\theta)$. Since F is invariant under signed permutations, it suffices to study

$$R(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Write $\kappa_{\mathrm{diag}} = \lambda_1 e_1^{\otimes d} + \lambda_2 e_2^{\otimes d}$ and denote

$$\Phi_{\mathrm{diag}}(\theta) = \sum_{i=1}^{2} \langle \kappa_{\mathrm{diag}}, q_i^{\otimes d} \rangle^2.$$

When $K = K_{\mathrm{diag}}$ (no PMI terms), one computes

$$d = 3: \quad \Phi_{\mathrm{diag}}(\theta) = (\lambda_1^2 + \lambda_2^2)(1 - 3\sin^2\theta\cos^2\theta),$$
$$d = 4: \quad \Phi_{\mathrm{diag}}(\theta) = (\lambda_1^2 + \lambda_2^2)(1 - 4u + 2u^2) + 4t_1 t_2 u^2, \qquad u = \sin^2\theta\cos^2\theta.$$

In both cases Φ_{diag} is globally maximized at $\theta = 0, \pi/2$, i.e. at $\mathrm{SP}(2)$. The drop away from $\mathrm{SP}(2)$ is quadratic in θ:

$$(14) \qquad\qquad d = 3: \quad \Phi_{\mathrm{diag}}(0) - \Phi_{\mathrm{diag}}(\theta) = \tfrac{3}{4}(\lambda_1^2 + \lambda_2^2)\sin^2(2\theta),$$

$$(15) \qquad\qquad d = 4: \quad \Phi_{\mathrm{diag}}(0) - \Phi_{\mathrm{diag}}(\theta) \geq \tfrac{1}{2}\,\mathrm{gap}_4^2\,\sin^2(2\theta),$$

where $\mathrm{gap}_4 = |\lambda_1 - \lambda_2|$. For arbitrary K_{off}, Cauchy–Schwarz gives

$$(16) \qquad F(R(\theta)) \;\leq\; \Phi_{\mathrm{diag}}(\theta) + 2\|\kappa_{\mathrm{off}}\|_{\mathrm{F}} \sum_{i=1}^{2} \left| \langle \kappa_{\mathrm{diag}}, q_i^{\otimes d} \rangle \right| + 2\|\kappa_{\mathrm{off}}\|_{\mathrm{F}}^2.$$

At $\theta = 0$ (any $P \in \mathrm{SP}(2)$), PMI zeros imply $F = \Phi_{\mathrm{diag}}(0) = \lambda_1^2 + \lambda_2^2$. For $\theta \neq 0$, the diagonal drop is quadratic in θ, while the perturbation from κ_{off} is linear. Thus for small θ, $\mathrm{SP}(2)$ are strict local maxima, but the global maximizer may drift away if $\|\kappa_{\mathrm{off}}\|_{\mathrm{F}}$ is not negligible.

This calculation illustrates the local versus global behavior under PMI. The ground-truth rotations $\mathrm{SP}(n)$ are locally stable: infinitesimal deviations reduce the objective. But the global landscape can change once off-diagonal PMI terms are present, creating new optima far from $\mathrm{SP}(n)$. This foreshadows the phenomena observed in our simulations.

8. Numerical experiments

In this section, we test the Riemannian gradient descent algorithm for PMICA introduced in Algorithm 1 and compare to ICA approaches. Consider the PMICA setup $\mathbf{x} = A\mathbf{s}$ with $A \in \mathrm{O}(n)$, where we assume that the data comes prewhitened. We estimate A by minimizing $F_N(Q) = \|\Pi_{V_{\mathrm{pmi}}^{\perp}}(Q^{\top} \bullet \widehat{\kappa}_4(\mathbf{x}))\|_{\mathrm{F}}^2$ over $\mathrm{O}(n)$ using RGD. We call the approach `RGD-PMICA`.

We compare to ICA, where we minimize $F_N^{\text{diag}}(Q) = \|\Pi_{V_{\text{diag}}^\perp}(Q^\top \bullet \widehat{\kappa}_4(\mathbf{x}))\|_F^2$ over $\mathrm{O}(n)$. We call this approach `RGD-ICA` when we minimize using RGD. We compare these two RGD approaches with classical ICA baselines `FastICA` [HO00] and `JADE` [CS93]. `FastICA` is a fixed-point algorithm that estimates independent components by maximizing non-Gaussianity, where orthogonality is maintained by whitening and a symmetric orthoognalization step at each iteration. `JADE` is a cumulant-based method that jointly diagonalizes fourth-order cumulant matrices via Jacobi (Givens) rotations, which enforce orthogonality by construction. As such, `JADE` is similar to `RGD-ICA` when $n = 2$.

Having estimated a mixing matrix $\widehat{A} \in \mathrm{O}(n)$, the ratio $\|\Pi_{V_{\text{diag}}^\perp}(\widehat{A}^\top \bullet \widehat{\kappa}_4(\mathbf{x}))\|_F/\|\widehat{\kappa}_4(\mathbf{x})\|_F$ measures the goodness of fit of ICA, and $\|\Pi_{V_{\text{pmi}}^\perp}(\widehat{A}^\top \bullet \widehat{\kappa}_4(\mathbf{x}))\|_F/\|\widehat{\kappa}_4(\mathbf{x})\|_F$ measures the goodness of fit of PMICA. We call these metrics "Distance to Independent" and "Distance to PMI", respectively. The denominator $\|\widehat{\kappa}_4(\mathbf{x})\|_F$ normalizes these metrics, ensuring that their values lie between zero and one.

8.1. Synthetic data. Let $\mathbf{z}^{(0)}$, and $\mathbf{z}^{(1)}$ be two-dimensional random vectors independent from each other with density functions

$$f_{\mathbf{z}^{(0)}}(z_1, z_2) = \frac{1}{2}|z_2|\mathbb{1}_{[-1,1]^2}(z_1, z_2), \qquad f_{\mathbf{z}^{(1)}}(z_1, z_2) = \frac{3}{2}|z_2|\mathbb{1}_{B_1}(z_1, z_2),$$

where $B_1 = \{(z_1, z_2) \in \mathbb{R}^2 \mid |z_1| + |z_2| \leq 1\}$ is the unit ℓ_1 ball and $\mathbb{1}_U$ is the indicator function for the set U. The vector $\mathbf{z}^{(0)}$ has independent entries and the entries of $\mathbf{z}^{(1)}$ are pairwise mean independent, which follows by direct calculations. For each $\alpha \in [0, 1]$, let $\mathbf{z}^{(\alpha)} = (1 - \alpha)\mathbf{z}^{(0)} + \alpha\mathbf{z}^{(1)}$ and let $\mathbf{s}^{(\alpha)}$ be the random vector obtained by rescaling each coordinate of $\mathbf{z}^{(\alpha)}$ to have unit variance. See Figure 2.
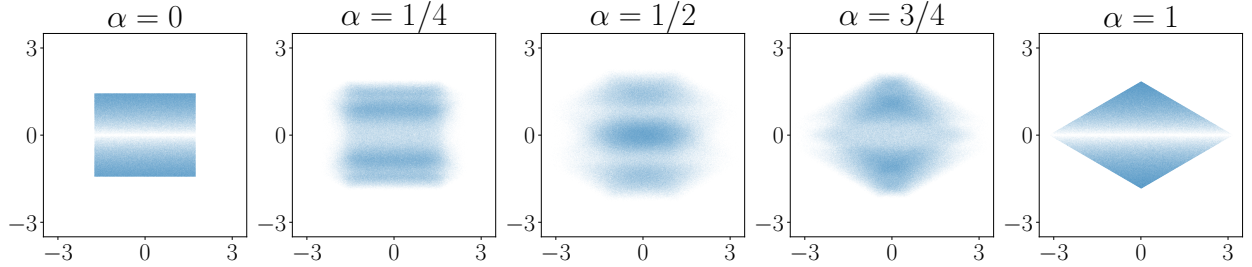

FIGURE 2. One million samples from $\mathbf{s}^{(\alpha)}$ for different α. It is pairwise mean independent for all α, because linear combinations of independent PMI vectors are PMI. It is independent when $\alpha = 0$ but not otherwise.

Given $A \in \mathrm{O}(2)$ and $\alpha \in [0, 1]$, consider the PMICA model $\mathbf{x}^{(\alpha)} = A\mathbf{s}^{(\alpha)}$. We investigate the performance of the different algorithms in recovering A given $N = 10^6$ samples from $\mathbf{x}^{(\alpha)}$, see Figure 3. In all plots in this section, each point is the median over 100 experiments. The fourth-order cumulant tensor $\kappa_4(\mathbf{s}^{(\alpha)}) \in S^4(\mathbb{R}^2)$ is generic in V_{pmi}, in the sense of Theorem 3.2, for all $\alpha \in [0, 1]$. Hence, an algorithm recovers the true mixing matrix A (up to signed permutation) if and only if the distance to PMI of $\widehat{A}^\top \bullet \widehat{\kappa}_4(\mathbf{x})$ is zero. Figure 3 shows that `RGD-PMICA` outperforms the ICA methods in recovering PMI distributions: ICA methods only recover the true sources when their distributions are close to independent (small α).

In Section 7, we quantified what close to independent means for ICA to recover the true rotation: $\mathrm{gap}_4(\widehat{\kappa}_4(\mathbf{s})) = |\widehat{\kappa}_4(\mathbf{s})_{1111} - \widehat{\kappa}_4(\mathbf{s})_{2222}|$ has to be a significant amount larger than the
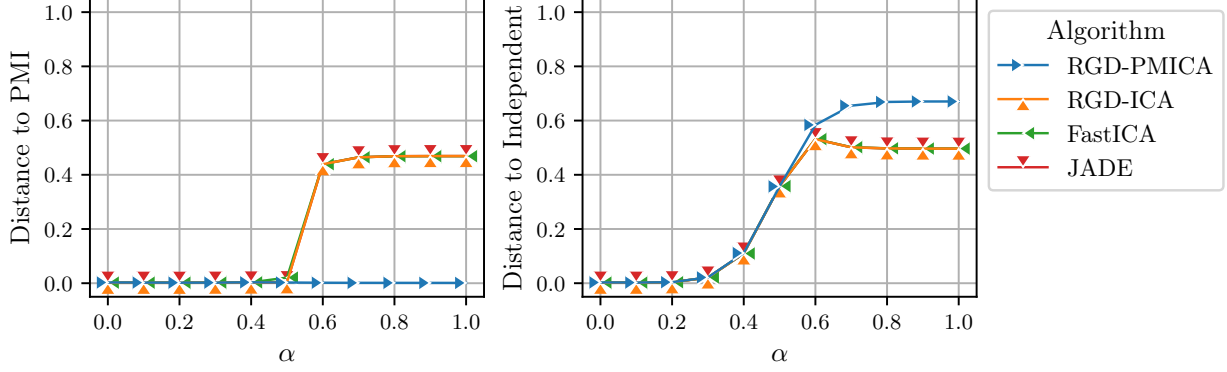
FIGURE 3. `RGD-PMICA` outperforms ICA algorithms in recovering $\mathbf{s}^{(\alpha)}$. The ICA methods find the closest independent distribution, which is not the PMI one for $\alpha \geq 0.6$.

off-diagonal Frobenius norm $\|\widehat{\kappa}_4(\mathbf{s})_{\mathrm{off}}\|_{\mathrm{F}} = \|\Pi_{V_{\mathrm{diag}}^\perp}(\widehat{\kappa}_4(\mathbf{s}))\|_{\mathrm{F}}$. Figure 4 shows these quantities for different $\mathbf{s}^{(\alpha)}$. When $\alpha = 0.6$, the ICA methods do not recover the true PMI source (as seen in Figure 3), and the ratio between gap_4 and $\|\widehat{\kappa}_4(\mathbf{s})_{\mathrm{off}}\|$ is 1.07, c.f. (15) and (16).
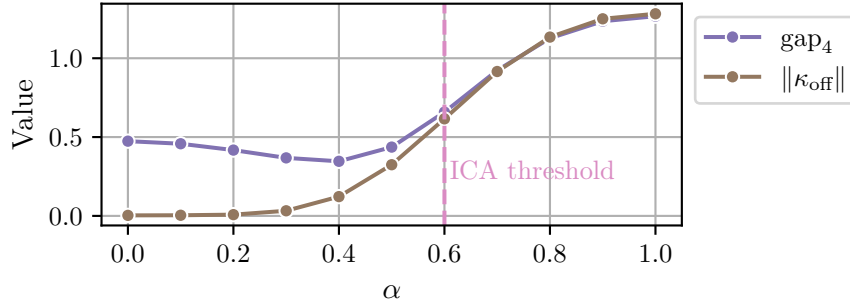
FIGURE 4. Gap ($|\widehat{\kappa}_{1111} - \widehat{\kappa}_{2222}|$) and off-diagonal Frobenius norm for $\widehat{\kappa}_4(\mathbf{s}^{(\alpha)})$. ICA models recover the PMI sources when α is smaller than the ICA threshold.

Now we study the performance of such algorithms in recovering PMI distributions for different dimensions. Consider the distribution with density

$$f(z_1, \ldots, z_n; \alpha) = \frac{1}{C(\alpha)} \prod_{i=1}^n |z_i|^{\alpha_i - 1} \mathbb{1}_{B_1^n}(z_1, \ldots, z_n),$$

where $C(\alpha)$ is a normalizing constant and $B_1^n = \{y \in \mathbb{R}^n \mid \|y\|_1 \leq 1\}$. This is closely related to the Dirichlet distribution. Let $\mathbf{s}$ be an n-dimensional random vector obtained from this distribution after rescaling each coordinate of $\mathbf{z}$ so that they have unit variance. Note that $\mathbf{s}$ is PMI for all n. In the following, we use $\alpha_i = 2^{\frac{i-1}{n-1}}$. We choose these values of α_i to ensure that $\kappa_4(\mathbf{s})$ is not close to V_{diag} and has sufficiently distinct diagonal entries. In particular, the fourth-order cumulant tensor $\kappa_4(\mathbf{s})$ is generic in V_{pmi} for all n, in the sense of Theorem 3.2. When $n = 2$ this distribution coincides with the one defined above for $\mathbf{s}^{(1)}$.

We evaluate the algorithms for different values of n using a sample size of 10^6. Figure 5 shows the best-case scenario for each method: the minimum after using each method with

25 random initializations. Again, `RGD-PMICA` outperforms the ICA algorithm in recovering PMI sources. In Figure 5, the distance to independent is higher than the distance to PMI. This is expected since $V_{\mathrm{diag}} \subsetneq V_{\mathrm{pmi}}$.
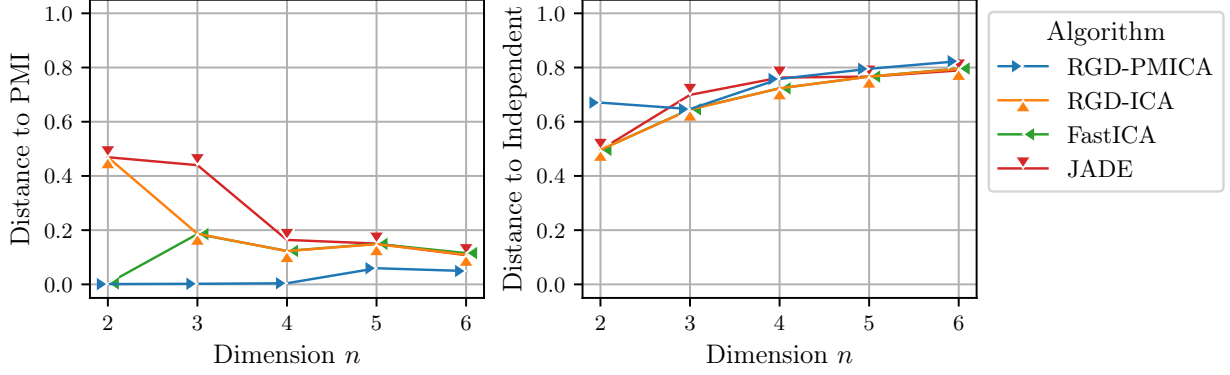

FIGURE 5. `RGD-PMICA` outperforms the ICA algorithms for all n. The decreasing trends of the ICA methods suggest that the closest independent distribution gets closer to the PMI one as n increases.

In the remainder of this section, we focus on `RGD-PMICA`. Figure 5 shows that `RGD-PMICA` fails to recover the true mixing matrix when $n \geq 5$ with 25 different initializations. This is because the complexity of the optimization problem increases as n grows. Figure 6 shows how the performance depends on the number of different random initializations. We observe that the number of initializations that `RGD-PMICA` requires to find the true PMI sources is of the order of 10^{n-2}, where n is the dimension.
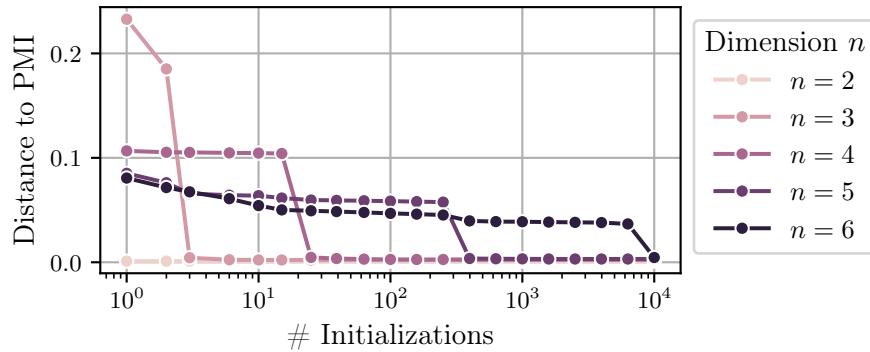

FIGURE 6. Convergence analysis of `RGD_PMICA`. The complexity of the optimization problem grows exponentially with the dimension.

Finally, we analyze the sample complexity to recover PMI sources. Figure 7 shows the standard deviation of $\widehat{\kappa}_4(\mathbf{s})$ one million samples over 100 experiments, and the distance to PMI obtained with `RGD-PMICA` using 10^4 initializations. We observe that the lines shared the same slope (provided the sample size is big enough). This slope is approximately $-1/2$ (in log-log scale), which agrees with the term $1/\sqrt{N}$ in Proposition 6.6.
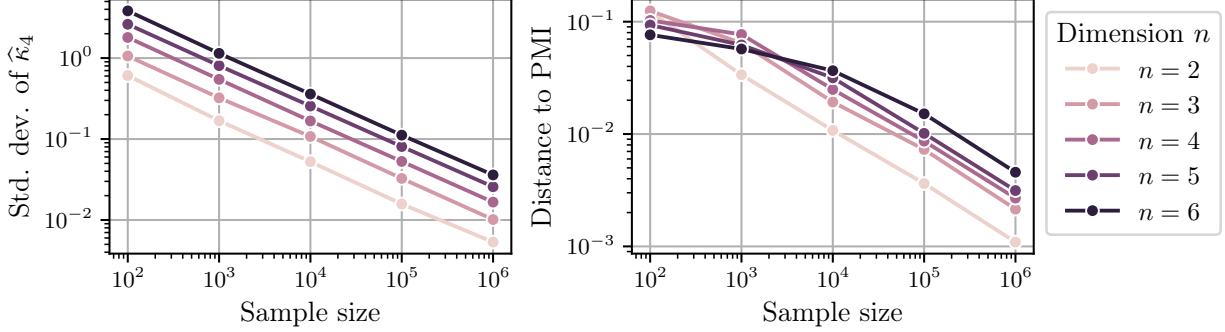
FIGURE 7. Sample complexity in different dimensions. Standard deviation of the estimator $\widehat{\kappa}_4(\mathbf{s})$ (left) and distance to PMI using `RGD-PMICA` (right). The lines corresponding to different dimensions share a common slope of $-1/2$.

8.2. Real data. In this section we test the performance of `RGD_PMICA` on a real world application. The electroencephalogram (EEG) dataset [CM10] records the brain's electrical activity of six subjects over two sessions. We use the data indexed by S01-1 in the 22nd dataset available at https://bnci-horizon-2020.eu/database/data-sets, which consists of 91648 observations of 64 electrodes. First, we perform dimensionality reduction via PCA. We keep the top 5 principal components, which explain 93% variability of the data. This leads to a 91648×5 data matrix. After whitening the data, we test the performance of the different approaches mentioned above to recover independent/PMI sources. See Table 2. `PCA` is the input data and `JADE` does not rely on any initialization. For each of the other optimization approaches and each metric, we report the best of 50 trials with different initial points in Table 2. Table 2 shows that `RGD-PMICA` outperforms the other methods in recovering PMI sources. It is interesting to observe that the distance to PMI achieved by `RGD-ICA` is approximately the one achieved by `RGD-PMICA`. This agrees with our discussion in Section 7. However, the rotation found by `RGD-ICA` minimizing the distance to independent (0.55) leads to a distance to PMI of 0.18, i.e., it does not correspond to the one that minimizes the distance to PMI (0.10).

Algorithm	Distance to PMI	Distance to Independent
PCA	0.50	0.97
RGD-PMICA	**0.09**	0.56
RGD-ICA	0.10	0.55
FastICA	0.41	0.69
JADE	0.34	**0.54**

TABLE 2. Goodness of fit of the PMICA and ICA models to EEG data

A typical qualitative measurement to analyze the performance of ICA approaches in EEG data is the ability to retrieve a source corresponding to the eye-blink artifact. Figure 8 shows that `RGD-PMICA` succeeds in finding the eye-blink artifact.
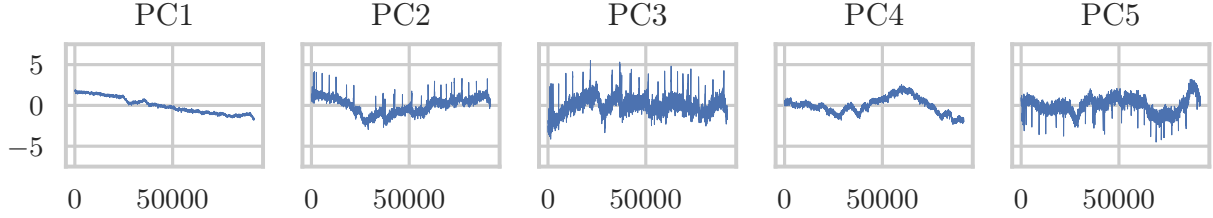
(A) Projection onto principal components
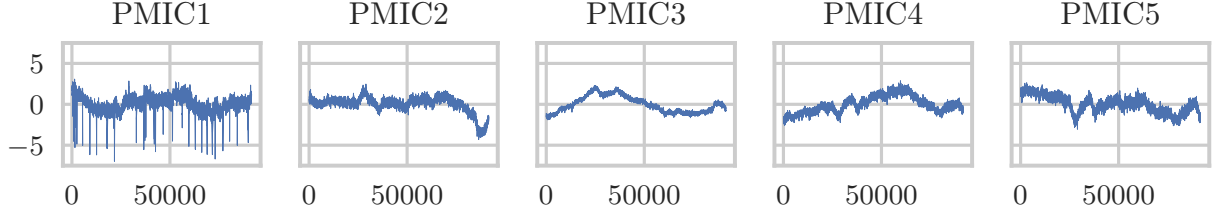
(B) Projection onto pairwise mean independent components

FIGURE 8. `RGD-PMICA` finds the eye-blink artifact (PMIC1)

Discussion

We established the identifiability of pairwise mean component analysis (PMICA), generalizing classical ICA by relaxing the independence assumption. It is more expressive than ICA, as it corresponds to a higher-dimensional family of cumulant tensors. Our results subsume previous examples of non-independent components analysis [MZ24] and cannot be further relaxed: if we drop one mean independence assumption on the source variables, the model becomes unidentifiable. We conclude by outlining directions for future work.

It is an open problem to determine the genericity conditions under which a symmetric tensor with an orthogonal basis of eigenvectors has a unique such basis. Our results in Section 4 address symmetric tensors of order up to nine.

We focused on genericity conditions for PMICA, fixing the order of the cumulant tensor. It is an open question to characterize distributional assumptions for genericity when all cumulants are available—the analog of the 'at most one Gaussian source' condition in ICA. It is also an open problem to extend to the case where $\mathbf{x}$ and $\mathbf{s}$ have different dimensions, following the study of full independence in the overcomplete setting [EK06, WS24].

Our concentration and parameter error bounds are universal in that they apply for any choice of V. We save a more refined analysis, particularly of improved moment and cumulant estimators for non-sub-Gaussian data, for future work.

Acknowledgments

References

[AGCSA25] Omar Al-Ghattas, Jiaheng Chen, and Daniel Sanz-Alonso. Sharp concentration of simple random tensors. *arXiv preprint arXiv:2502.16916*, 2025.

[AGH+14] Animashree Anandkumar, Rong Ge, Daniel J Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, 15(1):2773–2832, 2014.

[Akh20] Naum I. Akhiezer. *The Classical Moment Problem and Some Related Questions in Analysis*, volume 5 of *Classics in Applied Mathematics*. SIAM, Philadelphia, PA, 2020. Reprint of the 1965 English edition, originally published by Oliver and Boyd.

[AMS08] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[Bil95] Patrick Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, 3rd edition, 1995.

[BW97] Andrew D Back and Andreas S Weigend. A first application of independent component analysis to extracting structure from stock returns. *International journal of neural systems*, 8(04):473–484, 1997.

[Car98] J-F Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998.

[CGLM08] Pierre Comon, Gene Golub, Lek-Heng Lim, and Bernard Mourrain. Symmetric tensors and symmetric tensor rank. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1254–1279, 2008.

[CM10] Ricardo Chavarriaga and José del R Millán. Learning from EEG error-related potentials in noninvasive brain-computer interfaces. *IEEE transactions on neural systems and rehabilitation engineering*, 18(4):381–388, 2010.

[Com94] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[CS93] Jean-François Cardoso and Antoine Souloumiac. Blind beamforming for non-Gaussian signals. In *IEE proceedings F (radar and signal processing)*, volume 140, pages 362–370. IET, 1993.

[CS13] Dustin Cartwright and Bernd Sturmfels. The number of eigenvalues of a tensor. *Linear algebra and its applications*, 438(2):942–952, 2013.

[EK06] Jan Eriksson and Visa Koivunen. Complex random vectors and ICA models: Identifiability, uniqueness, and separability. *IEEE Transactions on Information theory*, 52(3):1017–1029, 2006.

[Fam70] Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2):383–417, May 1970.

[Fri13] Shmuel Friedland. Best rank one approximation of real symmetric tensors can be chosen symmetric. *Frontiers of Mathematics in China*, 8(1):19–40, 2013.

[GLS24] Marina Garrote-López and Monroe Stephenson. Cumulant tensors in partitioned independent component analysis. *arXiv preprint arXiv:2402.10089*, 2024.

[GS] Daniel R. Grayson and Michael E. Stillman. Macaulay2, a software system for research in algebraic geometry.

[Han82] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

[HHI01] Aapo Hyvärinen, Patrik O Hoyer, and Mika Inki. Topographic independent component analysis. *Neural computation*, 13(7):1527–1558, 2001.

[HO00] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[IR15] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, Cambridge, 2015.

[Jia25] Ziyu Jiang. Identification and estimation of simultaneous equation models using higher-order cumulant restrictions. *arXiv:2501.06777*, 2025.

[Kel70] 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.

[LCMS25] Paula Leyes Carreno, Chiara Meroni, and Anna Seigal. Linear causal disentanglement via higher-order cumulants. *La Matematica*, pages 1–40, 2025.

[Lim05] Lek-Heng Lim. Singular values and eigenvalues of tensors: a variational approach. In *1st IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, 2005.*, pages 129–132. IEEE, 2005.

[LZS20] CE Lee, X Zhang, and Xiaofeng Shao. Testing conditional mean independence for functional data. *Biometrika*, 107(2):331–346, 2020.

[Mar39] Józef Marcinkiewicz. Sur une propriété de la loi de Gauss. *Mathematische Zeitschrift*, 44(1):612–618, 1939.

[MBJS95] Scott Makeig, Anthony Bell, Tzyy-Ping Jung, and Terrence J Sejnowski. Independent component analysis of electroencephalographic data. *Advances in neural information processing systems*, 8, 1995.

[McC18] Peter McCullagh. *Tensor methods in statistics: Monographs on statistics and applied probability.* Chapman and Hall/CRC, 2018.

[MZ24] Geert Mesters and Piotr Zwiernik. Non-independent components analysis. *The Annals of Statistics*, 52(6):2506–2528, 2024.

[Qi11] Liqun Qi. The best rank-one approximation ratio of a tensor space. *SIAM Journal on matrix analysis and applications*, 32(2):430–442, 2011.

[Rob16] Elina Robeva. Orthogonal decomposition of symmetric tensors. *SIAM Journal on Matrix Analysis and Applications*, 37(1):86–102, 2016.

[Rom05] Steven Roman. *Advanced linear algebra*, volume 3. Springer, 2005.

[RSZ25] Alvaro Ribot, Anna Seigal, and Piotr Zwiernik. Orthogonal eigenvectors and singular vectors of tensors. *arXiv preprint arXiv:2506.19009*, 2025.

[RZ21] David Rossell and Piotr Zwiernik. Dependence in elliptical partial correlation graphs. *Electronic Journal of Statistics*, 15(2):4236–4263, 2021.

[Woo95] Jeffrey M Wooldridge. Selection corrections for panel data models under conditional mean independence assumptions. *Journal of econometrics*, 68(1):115–132, 1995.

[Woo10] Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data.* MIT press, 2010.

[WS24] Kexin Wang and Anna Seigal. Identifiability of overcomplete independent component analysis. *arXiv:2401.14709*, 2024.

[Zwi15] Piotr Zwiernik. *Semialgebraic statistics and latent tree models.* CRC Press, 2015.

Appendix A. Technical material from Section 6.3

A.1. The Gradient and Hessian of F, F_N. The functions F and F_N are defined in (8) and (10). The gradient $\mathrm{D}F(Q)$ of F at Q is the linear functional $\nabla F(Q)$: $\mathrm{D}F(Q)[\Delta] = \langle \nabla F(Q), \Delta \rangle$ on $\mathbb{R}^{n \times n}$. Let $\mathrm{grad}F$ denote its projection to the tangent space to $\mathrm{O}(n)$ at Q

$$T_Q\mathrm{O}(n) \;=\; \{\Delta \in \mathbb{R}^{n \times n} : Q^\top \Delta + \Delta^\top Q = 0\} = \{QA : \ A^\top = -A\}.$$

Fixing Q, we denote $U := T_Q\mathrm{O}(n)$. The Riemannian gradient of F at Q is

$$\mathrm{grad}F(Q) \;:=\; \Pi_U(\nabla F(Q)).$$

Since, for any $\Delta \in \mathbb{R}^{n \times n}$

$$\langle \mathrm{grad}F(Q), \Delta \rangle \;=\; \langle \mathrm{grad}F(Q), \Pi_U(\Delta) \rangle \;=\; \langle \nabla F(Q), \Pi_U(\Delta) \rangle$$

this restricts $\mathrm{D}F(Q)$ to $\Delta \in U$; see [AMS08, Chapter 3]. We express the gradients of F, F_N in terms of g, g_N.

Lemma A.1. *Let $Q \in \mathrm{O}(n)$ and let $U = T_Q\mathrm{O}(n)$ then*

$$\mathrm{grad}F(Q) = 2\Pi_U\Big((\mathrm{D}g(Q))^*[g(Q)]\Big) \quad and \quad \mathrm{grad}F_N(Q) = 2\Pi_U\Big((\mathrm{D}g_N(Q))^*[g_N(Q)]\Big),$$

where for a linear mapping G on $\mathbb{R}^{n \times n}$, G^* denotes its conjugate mapping.

PROOF. From the definition of the derivative $\mathrm{D}g(Q)$ of g at Q, we get

$$g(Q + \Delta) - g(Q) \;=\; \mathrm{D}g(Q)[\Delta] + o(\|\Delta\|).$$

It follows that

$$F(Q + \Delta) - F(Q) \;=\; 2\langle g(Q), \mathrm{D}g(Q)[\Delta] \rangle + o(\|\Delta\|) \;=\; 2\langle (\mathrm{D}g(Q))^*[g(Q)], \Delta \rangle + o(\|\Delta\|).$$

Hence the derivative $\mathrm{D}F(Q)$ is represented by the gradient $\nabla F(Q)$ which satisfies $\nabla F(Q) = 2\,(\mathrm{D}g(Q))^*[g(Q)] \in \mathbb{R}^{n \times n}$. The Riemannian gradient is its orthogonal projection to U. The argument for F_N is analogous. $\qquad\square$

Lemma A.2. *For any $Q \in \mathrm{O}(n)$,*

$$\|\mathrm{D}g_N(Q) - \mathrm{D}g(Q)\| \;\leq\; d\,\|\widehat{\mu}_d - \mu_d\|_{\mathrm{F}}, \qquad \|\mathrm{D}g(Q)\| \;\leq\; d\,\|\mu_d\|_{\mathrm{F}},$$

where $\| \cdot \|$ is the operator norm induced by matrix $\| \cdot \|_{\mathrm{F}}$ and tensor $\| \cdot \|_{\mathrm{F}}$.

PROOF. Let $\mathcal{T} := \widehat{\mu}_d - \mu_d$ and $H(Q) := Q^\top \bullet \mathcal{T}$ so that $\mathrm{D}g_N(Q) - \mathrm{D}g(Q) = \Pi_{V^\perp}(\mathrm{D}H(Q))$. For any $\Delta \in \mathbb{R}^{n \times n}$,

$$\mathrm{D}H(Q)[\Delta] = \sum_{j=1}^{d} \big(Q^\top\big)^{\otimes(j-1)} \otimes \Delta^\top \otimes \big(Q^\top\big)^{\otimes(d-j)} \;\bullet\; \mathcal{T}.$$

By multilinearity and $\|Q\| = 1$,

$$\|\mathrm{D}H(Q)[\Delta]\|_{\mathrm{F}} \;\leq\; \sum_{j=1}^{d} \|\Delta\|\,\|\mathcal{T}\|_{\mathrm{F}} \;\leq\; d\,\|\Delta\|_{\mathrm{F}}\,\|\mathcal{T}\|_{\mathrm{F}}.$$

Thus $\|\mathrm{D}H(Q)\| \leq d\,\|\mathcal{T}\|_{\mathrm{F}}$. Since $\Pi_{V^\perp}$ is an orthogonal projection, $\|\Pi_{V^\perp}\| = 1$, hence $\|\mathrm{D}g_N(Q) - \mathrm{D}g(Q)\| \leq d\,\|\mathcal{T}\|_{\mathrm{F}}$. The bound for $\mathrm{D}g(Q)$ follows by the same argument with $\mathcal{T}$ replaced by μ_d. $\qquad\square$

A.2. The spectral norm of a tensor. Let $\mathcal{T} \in S^d(\mathbb{R}^n)$. We collect useful results on the spectral norm

$$\|\mathcal{T}\|_2 = \max_{\|u\|=1} \langle \mathcal{T}, u^{\otimes d} \rangle,$$

defined in (11). One can view T as a multilinear form and define

$$\|\mathcal{T}\|_{2,\ldots,2} := \max_{\|u^{(1)}\|=\cdots=\|u^{(d)}\|=1} \langle \mathcal{T}, u^{(1)} \otimes \cdots \otimes u^{(d)} \rangle.$$

The norm $\|\mathcal{T}\|_{2,\ldots,2}$ (resp. $\|\mathcal{T}\|_2$) characterizes the best rank-one approximation (resp. symmetric approximation) of $\mathcal{T}$. For general tensors the two norms differ, but for symmetric tensors the best rank-one approximation may be chosen symmetric [Fri13, Theorem 1]. Hence

$$(17) \qquad \|\mathcal{T}\|_{2,\ldots,2} = \|\mathcal{T}\|_2$$

for all $\mathcal{T} \in S^d(\mathbb{R}^n)$. The next lemma compares the Frobenius norm and the spectral norm.

Lemma A.3. *For any* $\mathcal{T} \in S^d(\mathbb{R}^n)$, *we have* $\|\mathcal{T}\|_F \leq n^{(d-1)/2} \|\mathcal{T}\|_2$.

PROOF. For $d = 2, 3$ this appears in [Qi11, Theorem 4.3]. The author conjectured that the same bound holds for all $d \geq 4$ [Qi11, Conjecture 2]. As explained in the discussion following that conjecture, his Conjecture 1 together with [Qi11, Theorem 3.1] imply the result. Conjecture 1 is (17), which was established in [Fri13, Theorem 1]. $\square$

A.3. Proof of Proposition 6.7. By first-order optimality,

$$\operatorname{grad} F_N(\widehat{Q}) = 0, \qquad \operatorname{grad} F(Q) = 0 \quad \text{for every } Q \in Q^\star \mathrm{SP}(n).$$

Let $Q_0 \in Q^\star \mathrm{SP}(n)$ be the orbit point minimizing $\|\widehat{Q} - Q_0\|_F$, and let $\gamma : [0,1] \to \mathrm{O}(n)$ be the geodesic from Q_0 to $\widehat{Q}$ with tangent $\Delta = \dot{\gamma}(0) \in T_{Q_0}\mathrm{O}(n)$. A Taylor expansion of the empirical gradient along γ gives

$$0 = \operatorname{grad} F_N(\widehat{Q}) = \operatorname{grad} F_N(Q_0) + \operatorname{Hess} F_N(Q_0)[\Delta] + o(\|\Delta\|_F).$$

Since $\operatorname{grad} F(Q_0) = 0$, we obtain

$$\underbrace{\big(\operatorname{grad} F_N(Q_0) - \operatorname{grad} F(Q_0)\big)}_{\text{score fluctuation}} + \underbrace{\operatorname{Hess} F(Q_0)[\Delta]}_{\text{population curvature}} + \underbrace{\big(\operatorname{Hess} F_N(Q_0) - \operatorname{Hess} F(Q_0)\big)[\Delta]}_{\text{Hessian fluctuation}} + o(\|\Delta\|_F) = 0.$$

Both fluctuation terms are linear in $\mathcal{T} = \widehat{\mu}_d - \mu_d$. Indeed, at Q_0 we have $g(Q_0) = 0$, so, using Lemma A.1,

$$\operatorname{grad} F_N(Q_0) - \operatorname{grad} F(Q_0) = 2\Pi_U \Big(\big(Dg_N(Q_0)^*[g_N(Q_0)] - Dg(Q_0)^*[g(Q_0)] \big) \Big)$$
$$= 2\,\Pi_U \Big(Dg_N(Q_0)^*[g_N(Q_0) - g(Q_0)] \Big).$$

As before,

$$\|g_N(Q_0) - g(Q_0)\|_F = \|\Pi_{V^\perp}(Q_0^\top \bullet (\widehat{\mu}_d - \mu_d))\|_F \leq \|\widehat{\mu}_d - \mu_d\|_F.$$

Insert and subtract $Dg(Q_0)$:

$$Dg_N(Q_0)^*[g_N - g] = Dg(Q_0)^*[g_N - g] + \big(Dg_N(Q_0) - Dg(Q_0)\big)^*[g_N - g].$$

Taking Frobenius norms and using Lemma A.2 together with $\|g_N(Q_0) - g(Q_0)\|_F \leq \|\widehat{\mu}_d - \mu_d\|_F$, we obtain

$$\|\mathrm{grad}\, F_N(Q_0) - \mathrm{grad}\, F(Q_0)\|_F \leq 2\,\|Dg(Q_0)\|\,\|g_N(Q_0) - g(Q_0)\|_F$$
$$+ 2\,\|Dg_N(Q_0) - Dg(Q_0)\|\,\|g_N(Q_0) - g(Q_0)\|_F$$
$$\leq 2d\,\|\mu_d\|_F\,\|\widehat{\mu}_d - \mu_d\|_F + 2d\,\|\widehat{\mu}_d - \mu_d\|_F^2.$$

Similarly, differentiating $\mathrm{grad}\, F_N(Q) = 2\, Dg_N(Q)^* g_N(Q)$ and using $g(Q_0) = 0$ gives

$$\mathrm{Hess} F_N(Q_0)[\Delta] = 2\, Dg_N(Q_0)^* Dg_N(Q_0)[\Delta], \qquad \mathrm{Hess} F(Q_0)[\Delta] = 2\, Dg(Q_0)^* Dg(Q_0)[\Delta].$$

We write

$$\mathrm{Hess} F_N(Q_0) - \mathrm{Hess} F(Q_0) = 2\Big((Dg_N - Dg)^* Dg + Dg^*(Dg_N - Dg) + (Dg_N - Dg)^*(Dg_N - Dg)\Big),$$

all evaluated at Q_0. Hence, by Lemma A.2,

$$\|\mathrm{Hess} F_N(Q_0) - \mathrm{Hess} F(Q_0)\| \leq 4\|Dg(Q_0)\|\|Dg_N(Q_0) - Dg(Q_0)\| + 2\|Dg_N(Q_0) - Dg(Q_0)\|^2$$
$$\leq 4d^2\,\|\mu_d\|_F\,\|\widehat{\mu}_d - \mu_d\|_F + 2d^2\,\|\widehat{\mu}_d - \mu_d\|_F^2.$$

Thus, up to a quadratic remainder, both fluctuations are linear in $\|\widehat{\mu}_d - \mu_d\|_F$ with constants depending only on d and $\|\mu_d\|_F$. Invoking the curvature condition on $T_{Q_0}\mathrm{O}(n)$,

$$\langle \mathrm{Hess} F(Q_0)[\Delta], \Delta \rangle \geq \kappa\,\|\Delta\|_F^2,$$

and using the Taylor expansion above, we obtain (for $\|\widehat{\mu}_d - \mu_d\|_F$ small enough),

$$\kappa\,\|\Delta\|_F \lesssim \|\mathrm{grad}\, F_N(Q_0) - \mathrm{grad}\, F(Q_0)\|_F + \|\mathrm{Hess} F_N(Q_0) - \mathrm{Hess} F(Q_0)\|\,\|\Delta\|_F.$$

Absorbing the last term into the left-hand side and using the bounds above, we get

$$\|\Delta\|_F \lesssim \frac{d\,\|\mu_d\|_F}{\kappa}\,\|\widehat{\mu}_d - \mu_d\|_F + \frac{d}{\kappa}\,\|\widehat{\mu}_d - \mu_d\|_F^2.$$

The tangent space at $Q_0 \in \mathrm{O}(n)$ is $T_{Q_0}\mathrm{O}(n) = \{Q_0\Omega : \Omega^\top = -\Omega\}$. For any $\widehat{Q}$ sufficiently close to Q_0, there exists a unique skew-symmetric Ω such that

$$\widehat{Q} = Q_0 \exp(\Omega).$$

The geodesic $\gamma(t) = Q_0 \exp(t\Omega)$ then connects Q_0 to $\widehat{Q}$, with initial tangent $\Delta = \dot{\gamma}(0) = Q_0\Omega \in T_{Q_0}\mathrm{O}(n)$. Since multiplication by Q_0 is an isometry,

$$\|\Delta\|_F = \|\Omega\|_F.$$

Moreover,

$$\widehat{Q} - Q_0 = Q_0\big(\exp(\Omega) - I\big) = Q_0\big(\Omega + \tfrac{1}{2}\Omega^2 + \cdots\big),$$

so

$$\|\widehat{Q} - Q_0\|_F \leq \|\Omega\|_F + O(\|\Omega\|_F^2) = \|\Delta\|_F + O(\|\Delta\|_F^2).$$

Thus, for sufficiently small $\|\Delta\|_F$,

$$\|\widehat{Q} - Q_0\|_F \lesssim \|\Delta\|_F.$$

Combining with the bound on $\|\Delta\|_F$ obtained above yields

$$\|\widehat{Q} - Q_0\|_F \lesssim \frac{d\,\|\mu_d\|_F}{\kappa}\,\|\widehat{\mu}_d - \mu_d\|_F + \frac{d}{\kappa}\,\|\widehat{\mu}_d - \mu_d\|_F^2.$$

Since $Q_0 \in Q^\star\mathrm{SP}(n)$ was chosen to minimize $\|\widehat{Q} - Q\|_F$, the result follows.

The above argument shows that inequality (12) holds deterministically, given a realization of $\widehat{\mu}_d$. By Proposition 6.6, the deviation $\|\widehat{\mu}_d - \mu_d\|_2$ admits a sharp concentration bound with probability at least $1 - e^{-\tau}$. Using Lemma A.3 to relate Frobenius and spectral norms then yields the high-probability bound stated in Proposition 6.7.

ÁLVARO RIBOT, HARVARD UNIVERSITY
Email address: `aribotbarrado@g.harvard.edu`

ANNA SEIGAL, HARVARD UNIVERSITY
Email address: `aseigal@seas.harvard.edu`

PIOTR ZWIERNIK, UNIVERSITAT POMPEU FABRA AND BARCELONA SCHOOL OF ECONOMICS
Email address: `piotr.zwiernik@upf.edu`