Independent mechanism analysis, a new concept?

Luigi Gresele^{*1} **Julius von Kügelgen**^{*1,2} **Vincent Stimper** ^{1,2}

Bernhard Schölkopf ^{1} **Michel Besserve** ^{1}

^{1} Max Planck Institute for Intelligent Systems, Tübingen, Germany ^{2} University of Cambridge
{luigi.gresele,jvk,vincent.stimper,bs,besserve}@tue.mpg.de

Abstract

Independent component analysis provides a principled framework for unsupervised representation learning, with solid theory on the identifiability of the latent code that generated the data, given only observations of mixtures thereof. Unfortunately, when the mixing is nonlinear, the model is provably nonidentifiable, since statistical independence alone does not sufficiently constrain the problem. Identifiability can be recovered in settings where additional, typically observed variables are included in the generative process. We investigate an alternative path and consider instead including assumptions reflecting the principle of *independent causal mechanisms* exploited in the field of causality. Specifically, our approach is motivated by thinking of each source as independently influencing the mixing process. This gives rise to a framework which we term independent mechanism analysis. We provide theoretical and empirical evidence that our approach circumvents a number of nonidentifiability issues arising in nonlinear blind source separation.

1 Introduction

One of the goals of unsupervised learning is to uncover properties of the data generating process, such as latent structures giving rise to the observed data. Identifiability [55] formalises this desideratum: under suitable assumptions, a model learnt from observations should match the ground truth, up to well-defined ambiguities. Within representation learning, identifiability has been studied mostly in the context of independent component analysis (ICA) [17, 40], which assumes that the observed data $\mathbf{x}$ results from mixing unobserved *independent* random variables s_i referred to as *sources*. The aim is to recover the sources based on the observed mixtures alone, also termed *blind source separation* (BSS). A major obstacle to BSS is that, in the nonlinear case, independent component estimation does not necessarily correspond to recovering the *true* sources: it is possible to give counterexamples where the observations are transformed into components y_i which are independent, yet still mixed with respect to the true sources s_i [20, 39, 98]. In other words, nonlinear ICA is not identifiable.

In order to achieve identifiability, a growing body of research postulates additional supervision or structure in the data generating process, often in the form of *auxiliary variables* [28, 30, 37, 38, 41]. In the present work, we investigate a different route to identifiability by drawing inspiration from the field of *causal inference* [71, 78] which has provided useful insights for a number of machine learning tasks, including semi-supervised [87, 103], transfer [6, 23, 27, 31, 61, 72, 84, 85, 97, 102, 107], reinforcement [7, 14, 22, 26, 53, 59, 60, 106], and unsupervised [9, 10, 54, 70, 88, 91, 104, 105] learning. To this end, we *interpret the ICA mixing as a causal process* and apply the principle of independent causal mechanisms (ICM) which postulates that the generative process consists of independent modules which do not share information [43, 78, 87]. In this context, "independent" does not refer to *statistical* independence of random variables, but rather to the notion that the distributions and functions composing the generative process are chosen independently by Nature [43, 48]. While a formalisation of ICM [43, 57] in terms of algorithmic (Kolmogorov) complexity [51] exists, it is not computable, and hence applying ICM in practice requires assessing such non-statistical independence

[*] Equal contribution. Code available at: https://github.com/lgresele/independent-mechanism-analysis

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

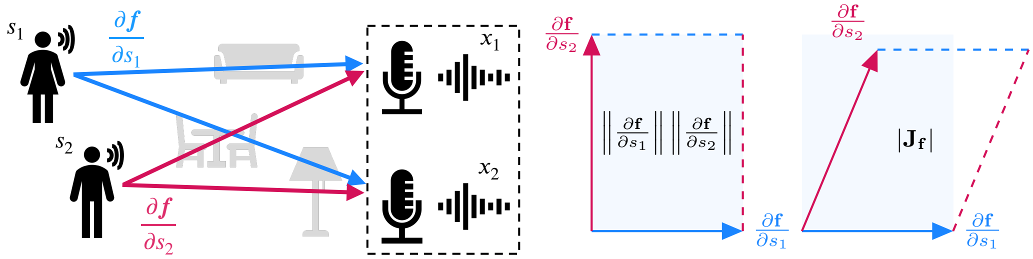

s_1
$\frac{\partial f}{\partial s_1}$
s_2
$\frac{\partial f}{\partial s_2}$
x_1
x_2
$\frac{\partial \mathbf{f}}{\partial s_2}$
$\left\|\frac{\partial \mathbf{f}}{\partial s_1}\right\| \left\|\frac{\partial \mathbf{f}}{\partial s_2}\right\|$
$\frac{\partial \mathbf{f}}{\partial s_1}$
$\frac{\partial \mathbf{f}}{\partial s_2}$
$|\mathbf{J_f}|$
$\frac{\partial \mathbf{f}}{\partial s_1}$

Figure 1: *(Left)* For the cocktail party problem, the ICM principle *as traditionally understood* would say that the content of speech $p_\mathbf{s}$ is independent of the mixing or recording process $\mathbf{f}$ (microphone placement, room acoustics). IMA refines, or extends, this idea *at the level of the mixing function* by postulating that the contributions $\partial\mathbf{f}/\partial s_i$ of each source to $\mathbf{f}$, as captured by the speakers' positions relative to the recording process, should not be fine-tuned to each other. *(Right)* We formalise this independence between the $\partial\mathbf{f}/\partial s_i$, which are the columns of the Jacobian $\mathbf{J_f}$, as an *orthogonality condition*: the absolute value of the determinant $|\mathbf{J_f}|$, i.e., the volume of the parallelepiped spanned by $\partial\mathbf{f}/\partial s_i$, should decompose as the product of the norms of the $\partial\mathbf{f}/\partial s_i$.

with suitable domain specific criteria [96]. The goal of our work is thus to *constrain the nonlinear ICA problem, in particular the mixing function, via suitable ICM measures*, thereby ruling out common counterexamples to identifiability which intuitively violate the ICM principle.

Traditionally, ICM criteria have been developed for causal discovery, where *both cause and effect are observed* [18, 45, 46, 110]. They enforce an independence between (i) the cause (source) distribution and (ii) the conditional or mechanism (mixing function) generating the effect (observations), and thus rely on the fact that the *observed* cause distribution is informative. As we will show, this renders them insufficient for nonlinear ICA, since the constraints they impose are satisfied by common counterexamples to identifiability. With this in mind, we introduce a new way to characterise or *refine* the ICM principle for unsupervised representation learning tasks such as nonlinear ICA.

Motivating example. To build intuition, we turn to a famous example of ICA and BSS: the cocktail party problem, illustrated in Fig. 1 *(Left)*. Here, a number of conversations are happening in parallel, and the task is to recover the individual voices s_i from the recorded mixtures x_i. The mixing or recording process $\mathbf{f}$ is primarily determined by the room acoustics and the locations at which microphones are placed. Moreover, each speaker influences the recording through their positioning in the room, and we may think of this influence as $\partial\mathbf{f}/\partial s_i$. Our independence postulate then amounts to stating that the speakers' positions are not fine-tuned to the room acoustics and microphone placement, or to each other, i.e., *the contributions $\partial\mathbf{f}/\partial s_i$ should be independent (in a non-statistical sense).*[1]

Our approach. We formalise this notion of independence between the contributions $\partial\mathbf{f}/\partial s_i$ of each source to the mixing process (i.e., the columns of the Jacobian matrix $\mathbf{J_f}$ of partial derivatives) as an orthogonality condition, see Fig. 1 *(Right)*. Specifically, the absolute value of the determinant $|\mathbf{J_f}|$, which describes the local change in infinitesimal volume induced by mixing the sources, should factorise or decompose as the product of the norms of its columns. This can be seen as a decoupling of the local influence of each partial derivative in the pushforward operation (mixing function) mapping the source distribution to the observed one, and gives rise to a novel framework which we term independent mechanism analysis (IMA). IMA can be understood as a refinement of the ICM principle that applies the idea of independence of mechanisms at the level of the mixing function.

Contributions. The structure and contributions of this paper can be summarised as follows:

- we review well-known obstacles to identifiability of nonlinear ICA (§ 2.1), as well as existing ICM criteria (§ 2.2), and show that the latter do not sufficiently constrain nonlinear ICA (§ 3);
- we propose a more suitable ICM criterion for unsupervised representation learning which gives rise to a new framework that we term independent mechanism analysis (IMA) (§ 4); we provide geometric and information-theoretic interpretations of IMA (§ 4.1), introduce an IMA contrast function which is invariant to the inherent ambiguities of nonlinear ICA (§ 4.2), and show that it rules out a large class of counterexamples and is consistent with existing identifiability results (§ 4.3);
- we experimentally validate our theoretical claims and propose a regularised maximum-likelihood learning approach based on the IMA constrast which outperforms the unregularised baseline (§ 5); additionally, we introduce a method to learn nonlinear ICA solutions with triangular Jacobian and a metric to assess BSS which can be of independent interest for the nonlinear ICA community.

[1] For additional intuition and possible violations in the context of the cocktail party problem, see Appendix B.4.

2 Background and preliminaries

Our work builds on and connects related literature from the fields of independent component analysis (§ 2.1) and causal inference (§ 2.2). We review the most important concepts below.

2.1 Independent component analysis (ICA)

Assume the following data-generating process for independent component analysis (ICA)

$$\mathbf{x} = \mathbf{f}(\mathbf{s}), \qquad\qquad p_{\mathbf{s}}(\mathbf{s}) = \textstyle\prod_{i=1}^{n} p_{s_i}(s_i), \tag{1}$$

where the *observed mixtures* $\mathbf{x} \in \mathbb{R}^n$ result from applying a *smooth and invertible mixing function* $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ to a set of *unobserved, independent signals or sources* $\mathbf{s} \in \mathbb{R}^n$ with smooth, factorised density $p_{\mathbf{s}}$ with connected support (see illustration Fig. 2b). The goal of ICA is to learn an *unmixing function* $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^n$ such that $\mathbf{y} = \mathbf{g}(\mathbf{x})$ has independent components. *Blind source separation* (BSS), on the other hand, aims to recover the true unmixing $\mathbf{f}^{-1}$ and thus the true sources $\mathbf{s}$ (up to tolerable ambiguities, see below). Whether performing ICA corresponds to solving BSS is related to the concept of *identifiability* of the model class. Intuitively, identifiability is the desirable property that *all models which give rise to the same mixture distribution should be "equivalent" up to certain ambiguities*, formally defined as follows.

Definition 2.1 ($\sim$-identifiability)**.** Let $\mathcal{F}$ be the set of all smooth, invertible functions $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$, and $\mathcal{P}$ be the set of all smooth, factorised densities $p_{\mathbf{s}}$ with connected support on $\mathbb{R}^n$. Let $\mathcal{M} \subseteq \mathcal{F} \times \mathcal{P}$ be a *subspace of models* and let $\sim$ be an *equivalence relation* on $\mathcal{M}$. Denote by $\mathbf{f}_* p_{\mathbf{s}}$ the *push-forward density* of $p_{\mathbf{s}}$ via $\mathbf{f}$. Then the generative process (1) is said to be *$\sim$-identifiable on $\mathcal{M}$* if

$$\forall (\mathbf{f}, p_{\mathbf{s}}), (\tilde{\mathbf{f}}, p_{\tilde{\mathbf{s}}}) \in \mathcal{M} : \qquad \mathbf{f}_* p_{\mathbf{s}} = \tilde{\mathbf{f}}_* p_{\tilde{\mathbf{s}}} \qquad \Longrightarrow \qquad (\mathbf{f}, p_{\mathbf{s}}) \sim (\tilde{\mathbf{f}}, p_{\tilde{\mathbf{s}}}) . \tag{2}$$

If the true model belongs to the model class $\mathcal{M}$, then $\sim$-identifiability ensures that any model in $\mathcal{M}$ learnt from (infinite amounts of) data will be $\sim$-equivalent to the true one. An example is *linear* ICA which is identifiable up to permutation and rescaling of the sources on the subspace $\mathcal{M}_{\mathrm{LIN}}$ of pairs of (i) invertible matrices (constraint on $\mathcal{F}$) and (ii) factorizing densities for which at most one s_i is Gaussian (constraint on $\mathcal{P}$) [17, 21, 93], see Appendix A for a more detailed account.

In the nonlinear case (i.e., without constraints on $\mathcal{F}$), identifiability is much more challenging. If s_i and s_j are independent, then so are $h_i(s_i)$ and $h_j(s_j)$ for any functions h_i and h_j. In addition to permutation-ambiguity, such *element-wise* $\mathbf{h}(\mathbf{s}) = (h_1(s_1), ..., h_n(s_n))$ can therefore not be resolved either. We thus define the desired form of identifiability for nonlinear BSS as follows.

Definition 2.2 ($\sim_{\mathrm{BSS}}$)**.** The equivalence relation $\sim_{\mathrm{BSS}}$ on $\mathcal{F} \times \mathcal{P}$ defined as in Defn. 2.1 is given by

$$(\mathbf{f}, p_{\mathbf{s}}) \sim_{\mathrm{BSS}} (\tilde{\mathbf{f}}, p_{\tilde{\mathbf{s}}}) \iff \exists \mathbf{P}, \mathbf{h} \quad \text{s.t.} \quad (\mathbf{f}, p_{\mathbf{s}}) = (\tilde{\mathbf{f}} \circ \mathbf{h}^{-1} \circ \mathbf{P}^{-1}, (\mathbf{P} \circ \mathbf{h})_* p_{\tilde{\mathbf{s}}}) \tag{3}$$

where $\mathbf{P}$ is a permutation and $\mathbf{h}(\mathbf{s}) = (h_1(s_1), ..., h_n(s_n))$ is an invertible, element-wise function.

A fundamental obstacle—and a crucial difference to the linear problem—is that in the nonlinear case, different mixtures of s_i and s_j can be independent, i.e., solving ICA is *not* equivalent to solving BSS. A prominent example of this is given by the *Darmois construction* [20, 39].

Definition 2.3 (Darmois construction)**.** The *Darmois construction* $\mathbf{g}^{\mathrm{D}} : \mathbb{R}^n \to (0,1)^n$ is obtained by recursively applying the conditional cumulative distribution function (CDF) transform:

$$g_i^{\mathrm{D}}(\mathbf{x}_{1:i}) := \mathbb{P}(X_i \leq x_i | \mathbf{x}_{1:i-1}) = \textstyle\int_{-\infty}^{x_i} p(x_i' | \mathbf{x}_{1:i-1}) dx_i' \qquad\qquad (i = 1, ..., n). \tag{4}$$

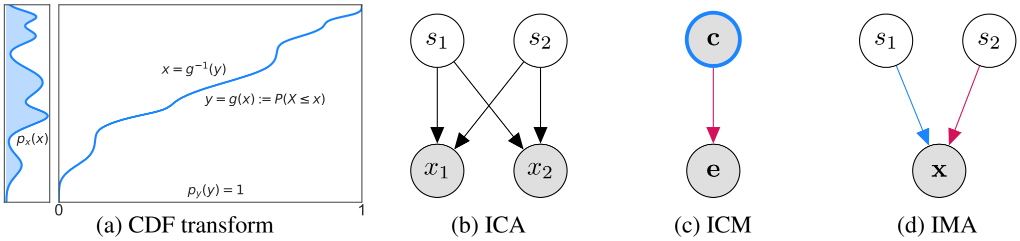
The resulting *estimated* sources $\mathbf{y}^{\mathrm{D}} = \mathbf{g}^{\mathrm{D}}(\mathbf{x})$ are mutually-independent uniform r.v.s by construction, see Fig. 2a for an illustration. However, they need not be meaningfully related to the *true* sources $\mathbf{s}$, and will, in general, still be a nonlinear mixing thereof [39].[2] Denoting the mixing function corresponding to (4) by $\mathbf{f}^{\mathrm{D}} = (\mathbf{g}^{\mathrm{D}})^{-1}$ and the uniform density on $(0,1)^n$ by $p_{\mathbf{u}}$, the *Darmois solution* $(\mathbf{f}^{\mathrm{D}}, p_{\mathbf{u}})$ thus allows construction of counterexamples to $\sim_{\mathrm{BSS}}$-identifiability on $\mathcal{F} \times \mathcal{P}$.[3]

Remark 2.4. $\mathbf{g}^{\mathrm{D}}$ has lower-triangular Jacobian, i.e., ${\partial g_i^{\mathrm{D}}}/{\partial x_j} = 0$ for $i < j$. Since the order of the x_i is arbitrary, applying $\mathbf{g}^{\mathrm{D}}$ after a permutation yields a different Darmois solution. Moreover, (4) yields independent components $\mathbf{y}^{\mathrm{D}}$ even if the sources s_i were not independent to begin with.[4]

[2] Consider, e.g., a mixing $\mathbf{f}$ with full Jacobian which yields a contradiction to Defn. 2.2, due to Remark 2.4.

[3] By applying a change of variables, we can see that the transformed variables in (4) are uniformly distributed in the open unit cube, thereby corresponding to independent components [69, § 2.2].

[4] This has broad implications for unsupervised learning, as it shows that, for i.i.d. observations, not only factorised priors, but *any* unconditional prior is insufficient for identifiability (see, e.g., [49], Appendix D.2).



(a) CDF transform (b) ICA (c) ICM (d) IMA

Figure 2: (a) Any observed density p_x can be mapped to a uniform p_y via the CDF transform $g(x) = \mathbb{P}(X \leq x)$; Darmois solutions $(\mathbf{f}^{\mathrm{D}}, p_{\mathbf{u}})$ constructed from (4) therefore automatically satisfy the independence postulated by IGCI (6). (b) ICA setting with $n = 2$ sources (shaded nodes are observed, white ones are unobserved). (c) Existing ICM criteria typically enforce independence between an observed input or cause distribution $p_{\mathbf{c}}$ and a mechanism $p_{\mathbf{e}|\mathbf{c}}$ (independent objects are highlighted in blue and red). (d) IMA enforces independence between the contributions of different sources s_i to the mixing function $\mathbf{f}$ as captured by $^{\partial \mathbf{f}}/_{\partial s_i}$.

Another well-known obstacle to identifiability are *measure-preserving automorphisms* (MPAs) of the source distribution $p_{\mathbf{s}}$: these are functions $\mathbf{a}$ which map the source space to itself without affecting its distribution, i.e., $\mathbf{a}_* p_{\mathbf{s}} = p_{\mathbf{s}}$ [39]. A particularly instructive class of MPAs is the following [49, 58].

Definition 2.5 ("Rotated-Gaussian" MPA)**.** Let $\mathbf{R} \in O(n)$ be an orthogonal matrix, and denote by $\mathbf{F}_{\mathbf{s}}(\mathbf{s}) = (F_{s_1}(s_1), ..., F_{s_n}(s_n))$ and $\mathbf{\Phi}(\mathbf{z}) = (\Phi(z_1), ..., \Phi(z_n))$ the element-wise CDFs of a smooth, factorised density $p_{\mathbf{s}}$ and of a Gaussian, respectively. Then the "rotated-Gaussian" MPA $\mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}})$ is

$$\mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}}) = \mathbf{F}_{\mathbf{s}}^{-1} \circ \mathbf{\Phi} \circ \mathbf{R} \circ \mathbf{\Phi}^{-1} \circ \mathbf{F}_{\mathbf{s}} \, . \tag{5}$$

$\mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}})$ first maps to the (rotationally invariant) standard isotropic Gaussian (via $\mathbf{\Phi}^{-1} \circ \mathbf{F}_{\mathbf{s}}$), then applies a rotation, and finally maps back, without affecting the distribution of the estimated sources. Hence, if $(\tilde{\mathbf{f}}, p_{\tilde{\mathbf{s}}})$ is a valid solution, then so is $(\tilde{\mathbf{f}} \circ \mathbf{a}^{\mathbf{R}}(p_{\tilde{\mathbf{s}}}), p_{\tilde{\mathbf{s}}})$ for any $\mathbf{R} \in O(n)$. Unless $\mathbf{R}$ is a permutation, this constitutes another common counterexample to $\sim_{\mathrm{BSS}}$-identifiability on $\mathcal{F} \times \mathcal{P}$.

Identifiability results for nonlinear ICA have recently been established for settings where an auxiliary variable $\mathbf{u}$ (e.g., environment index, time stamp, class label) renders the sources *conditionally* independent [37, 38, 41, 49]. The assumption on $p_{\mathbf{s}}$ in (1) is replaced with $p_{\mathbf{s}|\mathbf{u}}(\mathbf{s}|\mathbf{u}) = \prod_{i=1}^{n} p_{s_i|\mathbf{u}}(s_i|\mathbf{u})$, thus restricting $\mathcal{P}$ in Defn. 2.1. In most cases, $\mathbf{u}$ is assumed to be observed, though [30] is a notable exception. Similar results exist given access to a second noisy view $\tilde{\mathbf{x}}$ [28].

2.2 Causal inference and the principle of independent causal mechanisms (ICM)

Rather than relying only on additional assumptions on $\mathcal{P}$ (e.g., via auxiliary variables), we seek to further constrain (1) by also placing assumptions on the set $\mathcal{F}$ of mixing functions $\mathbf{f}$. To this end, we draw inspiration from the field of causal inference [71, 78]. Of central importance to our approach is the *Principle of Independent Causal Mechanisms* (ICM) [43, 56, 87].

Principle 2.6 (ICM principle [78])**.** *The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.*

These "modules" are typically thought of as the conditional distributions of each variable given its direct causes. Intuitively, the principle then states that these *causal conditionals* correspond to *independent mechanisms of nature* which do not share information. Crucially, here "independent" does not refer to *statistical* independence of random variables, but rather to independence of the underlying distributions as *algorithmic* objects. For a bivariate system comprising a cause $\mathbf{c}$ and an effect $\mathbf{e}$, this idea reduces to an independence of cause and mechanism, see Fig. 2c. One way to formalise ICM uses Kolmogorov complexity $K(\cdot)$ [51] as a measure of algorithmic information [43].

However, since Kolmogorov complexity is is not computable, using ICM in practice requires assessing Principle 2.6 with other suitable proxy criteria [9, 11, 34, 42, 45, 65, 75–78, 90, 110].[5] Allowing for deterministic relations between cause (sources) and effect (observations), the criterion which is most closely related to the ICA setting in (1) is *information-geometric causal inference* (IGCI) [18, 46].[6] IGCI assumes a nonlinear relation $\mathbf{e} = \mathbf{f}(\mathbf{c})$ and formulates a notion of indepen-

[5] "This can be seen as an algorithmic analog of replacing the empirically undecidable question of statistical independence with practical independence tests that are based on assumptions on the underlying distribution" [43].

[6] For a similar criterion which assumes linearity [45, 110] and its relation to linear ICA, see Appendix B.1.

dence between the cause distribution $p_{\mathbf{c}}$ and the deterministic mechanism $\mathbf{f}$ (which we think of as a degenerate conditional $p_{\mathbf{e}|\mathbf{c}}$) via the following condition (in practice, assumed to hold approximately),

$$C_{\text{IGCI}}(\mathbf{f}, p_{\mathbf{c}}) := \int \log |\mathbf{J}_{\mathbf{f}}(\mathbf{c})| \, p_{\mathbf{c}}(\mathbf{c}) d\mathbf{c} - \int \log |\mathbf{J}_{\mathbf{f}}(\mathbf{c})| \, d\mathbf{c} = 0 \,, \tag{6}$$

where $(\mathbf{J}_{\mathbf{f}}(\mathbf{c}))_{ij} = {}^{\partial f_i}/_{\partial c_j}(\mathbf{c})$ is the Jacobian matrix and $|\cdot|$ the absolute value of the determinant. C_{IGCI} can be understood as the covariance between $p_{\mathbf{c}}$ and $\log |\mathbf{J}_{\mathbf{f}}|$ (viewed as r.v.s on the unit cube w.r.t. the Lebesgue measure), so that $C_{\text{IGCI}} = 0$ rules out a form of fine-tuning between $p_{\mathbf{c}}$ and $|\mathbf{J}_{\mathbf{f}}|$. As its name suggests, IGCI can, from an information-geometric perspective, also be seen as an orthogonality condition between cause and mechanism in the space of probability distributions [46], see Appendix B.2, particularly eq. (19) for further details.

3 Existing ICM measures are insufficient for nonlinear ICA

Our aim is to use the ICM Principle 2.6 to further constrain the space of models $\mathcal{M} \subseteq \mathcal{F} \times \mathcal{P}$ and rule out common counterexamples to identifiability such as those presented in § 2.1. Intuitively, both the Darmois construction (4) and the rotated Gaussian MPA (5) give rise to "*non-generic*" solutions which should violate ICM: the former, $(\mathbf{f}^{\text{D}}, p_{\mathbf{u}})$, due the triangular Jacobian of $\mathbf{f}^{\text{D}}$ (see Remark 2.4), meaning that each observation $x_i = f_i^{\text{D}}(\mathbf{y}_{1:i})$ only depends on a subset of the inferred independent components $\mathbf{y}_{1:i}$, and the latter, $(\mathbf{f} \circ \mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}}), p_{\mathbf{s}})$, due to the dependence of $\mathbf{f} \circ \mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}})$ on $p_{\mathbf{s}}$ (5).

However, the ICM criteria described in § 2.2 were developed for the task of cause-effect inference where *both variables are observed*. In contrast, in this work, we consider an unsupervised representation learning task where *only the effects* (mixtures $\mathbf{x}$) *are observed*, but the causes (sources $\mathbf{s}$) are not. It turns out that this renders existing ICM criteria insufficient for BSS: they can easily be satisfied by spurious solutions which are not equivalent to the true one. We can show this for IGCI. Denote by $\mathcal{M}_{\text{IGCI}} = \{(\mathbf{f}, p_{\mathbf{s}}) \in \mathcal{F} \times \mathcal{P} : C_{\text{IGCI}}(\mathbf{f}, p_{\mathbf{s}}) = 0\} \subset \mathcal{F} \times \mathcal{P}$ the class of nonlinear ICA models satisfying IGCI (6). Then the following negative result holds.

Proposition 3.1 (IGCI is insufficient for $\sim_{\text{BSS}}$-identifiability)**.** *(1) is not $\sim_{\text{BSS}}$-identifiable on $\mathcal{M}_{\text{IGCI}}$.*

Proof. IGCI (6) is satisfied when $p_{\mathbf{s}}$ is uniform. However, the Darmois construction (4) yields uniform sources, see Fig. 2a. This means that $(\mathbf{f}^{\text{D}} \circ \mathbf{a}^{\mathbf{R}}(p_{\mathbf{u}}), p_{\mathbf{u}}) \in \mathcal{M}_{\text{IGCI}}$, so IGCI can be satisfied by solutions which do not separate the sources in the sense of Defn. 2.2, see footnote 2 and [39]. □

As illustrated in Fig. 2c, condition (6) and other similar criteria enforce a notion of "genericity" or "decoupling" of the mechanism w.r.t. the *observed* input distribution.[7] They thus rely on the fact that the cause (source) distribution is informative, and are generally not invariant to reparametrisation of the cause variables. In the (nonlinear) ICA setting, on the other hand, the *learnt* source distribution may be fairly uninformative. This poses a challenge for existing ICM criteria since any mechanism is generic w.r.t. an uninformative (uniform) input distribution.

4 Independent mechanism analysis (IMA)

As argued in § 3, enforcing independence between the input distribution and the mechanism (Fig. 2c), as existing ICM criteria do, is insufficient for ruling out spurious solutions to nonlinear ICA. We therefore propose a new ICM-inspired framework which is more suitable for BSS and which we term *independent mechanism analysis* (IMA).[8] All proofs are provided in Appendix C.

4.1 Intuition behind IMA

As motivated using the cocktail party example in § 1 and Fig. 1 *(Left)*, our main idea is to enforce a notion of *independence between the contributions or influences of the different sources s_i on the observations* $\mathbf{x} = \mathbf{f}(\mathbf{s})$ as illustrated in Fig. 2d—as opposed to between the source distribution and mixing function, cf. Fig. 2c. These contributions or influences are captured by the vectors of partial derivatives ${}^{\partial \mathbf{f}}/_{\partial s_i}$. IMA can thus be understood as a *refinement of ICM at the level of the mixing* $\mathbf{f}$: in addition to *statistically independent components* s_i, we look for a mixing with *contributions* ${}^{\partial \mathbf{f}}/_{\partial s_i}$ *which are independent*, in a non-statistical sense which we formalise as follows.

Principle 4.1 (IMA)**.** *The mechanisms by which each source s_i influences the observed distribution, as captured by the partial derivatives ${}^{\partial \mathbf{f}}/_{\partial s_i}$, are independent of each other in the sense that for all* $\mathbf{s}$*:*

$$\log |\mathbf{J}_{\mathbf{f}}(\mathbf{s})| = \sum_{i=1}^{n} \log \left\| \frac{\partial \mathbf{f}}{\partial s_i}(\mathbf{s}) \right\| \tag{7}$$

[7] In fact, many ICM criteria can be phrased as special cases of a unifying group-invariance framework [9].

[8] The title of the present work is thus a reverence to Pierre Comon's seminal 1994 paper [17].

Geometric interpretation. Geometrically, the IMA principle can be understood as an *orthogonality condition*, as illustrated for $n = 2$ in Fig. 1 *(Right)*. First, the vectors of partial derivatives ${}^{\partial \mathbf{f}}/_{\partial s_i}$, for which the IMA principle postulates independence, are the *columns* of $\mathbf{J_f}$. $|\mathbf{J_f}|$ thus measures the volume of the $n-$dimensional parallelepiped spanned by these columns, as shown on the right. The product of their norms, on the other hand, corresponds to the volume of an n-dimensional box, or rectangular parallelepiped with side lengths $\|{}^{\partial \mathbf{f}}/_{\partial s_i}\|$, as shown on the left. The two volumes are equal if and only if all columns ${}^{\partial \mathbf{f}}/_{\partial s_i}$ of $\mathbf{J_f}$ are orthogonal. Note that (7) is trivially satisfied for $n = 1$, i.e., if there is no mixing, further highlighting its difference from ICM for causal discovery.

Independent influences and orthogonality. In a high dimensional setting (large n), this orthogonality can be intuitively interpreted from the ICM perspective as *Nature choosing the direction of the influence of each source component in the observation space independently and from an isotropic prior*. Indeed, it can be shown that the scalar product of two independent isotropic random vectors in $\mathbb{R}^n$ vanishes as the dimensionality n increases (equivalently: two high-dimensional isotropic vectors are typically orthogonal). This property was previously exploited in other linear ICM-based criteria (see [44, Lemma 5] and [45, Lemma 1 & Thm. 1]).[9] The principle in (7) can be seen as a constraint on the function space, enforcing such orthogonality between the columns of the Jacobian of $\mathbf{f}$ at all points in the source domain, thus approximating the high-dimensional behavior described above.[10]

Information-geometric interpretation and comparison to IGCI. The additive contribution of the sources' influences ${}^{\partial \mathbf{f}}/_{\partial s_i}$ in (7) suggests their local *decoupling at the level of the mechanism* $\mathbf{f}$. Note that IGCI (6), on the other hand, postulates a different type of decoupling: one between $\log |\mathbf{J_f}|$ and $p_\mathbf{s}$. There, dependence between cause and mechanism can be conceived as a fine tuning between the derivative of the mechanism and the input density. The IMA principle leads to a complementary, non-statistical measure of independence between the influences ${}^{\partial \mathbf{f}}/_{\partial s_i}$ of the individual sources on the vector of observations. Both the IGCI and IMA postulates have an information-geometric interpretation related to the influence of ("non-statistically") independent modules on the observations: both lead to an *additive decomposition of a KL-divergence between the effect distribution and a reference distribution.* For IGCI, independent modules correspond to the cause distribution and the mechanism mapping the cause to the effect (see (19) in Appendix B.2). For IMA, on the other hand, these are the influences of each source component on the observations in an interventional setting (under soft interventions on individual sources), as measured by the KL-divergences between the original and intervened distributions. See Appendix B.3, and especially (22), for a more detailed account.

We finally remark that while recent work based on the ICM principle has mostly used the term "mechanism" to refer to causal Markov kernels $p(X_i|PA_i)$ or structural equations [78], we employ it in line with the broader use of this concept in the philosophical literature.[11] To highlight just two examples, [86] states that *"Causal processes, causal interactions, and causal laws provide the mechanisms by which the world works; to understand why certain things happen, we need to see how they are produced by these mechanisms"*; and [99] states that *"Mechanisms are events that alter relations among some specified set of elements"*. Following this perspective, we argue that a causal mechanism can more generally denote any process that describes the way in which causes influence their effects: the partial derivative ${}^{\partial \mathbf{f}}/_{\partial s_i}$ thus reflects a causal mechanism in the sense that it describes the infinitesimal changes in the observations $\mathbf{x}$, when an infinitesimal perturbation is applied to s_i.

4.2 Definition and useful properties of the IMA contrast

We now introduce a contrast function based on the IMA principle (7) and show that it possesses several desirable properties in the context of nonlinear ICA. First, we define a local contrast as the difference between the two integrands of (7) for a particular value of the sources $\mathbf{s}$.

Definition 4.2 (Local IMA contrast)**.** The local IMA contrast $c_{\text{IMA}}(\mathbf{f}, \mathbf{s})$ of $\mathbf{f}$ at a point $\mathbf{s}$ is given by

$$c_{\text{IMA}}(\mathbf{f}, \mathbf{s}) = \sum_{i=1}^{n} \log \left\| \frac{\partial \mathbf{f}}{\partial s_i}(\mathbf{s}) \right\| - \log |\mathbf{J_f}(\mathbf{s})| \ . \tag{8}$$

Remark 4.3. This corresponds to the left KL measure of diagonality [2] for $\sqrt{\mathbf{J_f}(\mathbf{s})^\top \mathbf{J_f}(\mathbf{s})}$.

[9] This has also been used as a *"leading intuition"* [sic] to interpret IGCI in [46].

[10] To provide additional intuition on how IMA differs from existing principles of independence of cause and mechanism, we give examples, both technical and pictorial, of violations of both in Appendix B.4.

[11] See Table 1 in [62] for a long list of definitions from the literature.

The local IMA contrast $c_{\text{IMA}}(\mathbf{f}, \mathbf{s})$ quantifies the extent to which the IMA principle is violated at a given point $\mathbf{s}$. We summarise some of its properties in the following proposition.

Proposition 4.4 (Properties of $c_{\text{IMA}}(\mathbf{f}, \mathbf{s})$)**.** *The local IMA contrast* $c_{\text{IMA}}(\mathbf{f}, \mathbf{s})$ *defined in* (8) *satisfies:*

(i) $c_{\text{IMA}}(\mathbf{f}, \mathbf{s}) \geq 0$*, with equality if and only if all columns* ${}^{\partial \mathbf{f}}/_{\partial s_i}(\mathbf{s})$ *of* $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ *are orthogonal.*

(ii) $c_{\text{IMA}}(\mathbf{f}, \mathbf{s})$ *is invariant to left multiplication of* $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ *by an orthogonal matrix and to right multiplication by permutation and diagonal matrices.*

Property *(i)* formalises the geometric interpretation of IMA as an orthogonality condition on the columns of the Jacobian from § 4.1, and property *(ii)* intuitively states that changes of orthonormal basis and permutations or rescalings of the columns of $\mathbf{J}_{\mathbf{f}}$ do not affect their orthogonality. Next, we define a global IMA contrast w.r.t. a source distribution $p_{\mathbf{s}}$ as the expected local IMA contrast.

Definition 4.5 (Global IMA contrast)**.** The global IMA contrast $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}})$ of $\mathbf{f}$ w.r.t. $p_{\mathbf{s}}$ is given by

$$C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}}) = \mathbb{E}_{\mathbf{s} \sim p_{\mathbf{s}}}[c_{\text{IMA}}(\mathbf{f}, \mathbf{s})] = \int c_{\text{IMA}}(\mathbf{f}, \mathbf{s}) p_{\mathbf{s}}(\mathbf{s}) d\mathbf{s}\,. \tag{9}$$

The global IMA contrast $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}})$ thus quantifies the extent to which the IMA principle is violated for a particular solution $(\mathbf{f}, p_{\mathbf{s}})$ to the nonlinear ICA problem. We summarise its properties as follows.

Proposition 4.6 (Properties of $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}})$)**.** *The global IMA contrast* $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}})$ *from* (9) *satisfies:*

(i) $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}}) \geq 0$*, with equality iff.* $\mathbf{J}_{\mathbf{f}}(\mathbf{s}) = \mathbf{O}(\mathbf{s})\mathbf{D}(\mathbf{s})$ *almost surely w.r.t.* $p_{\mathbf{s}}$*, where* $\mathbf{O}(\mathbf{s}), \mathbf{D}(\mathbf{s}) \in \mathbb{R}^{n \times n}$ *are orthogonal and diagonal matrices, respectively;*

(ii) $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}}) = C_{\text{IMA}}(\tilde{\mathbf{f}}, p_{\tilde{\mathbf{s}}})$ *for any* $\tilde{\mathbf{f}} = \mathbf{f} \circ \mathbf{h}^{-1} \circ \mathbf{P}^{-1}$ *and* $\tilde{\mathbf{s}} = \mathbf{Ph}(\mathbf{s})$*, where* $\mathbf{P} \in \mathbb{R}^{n \times n}$ *is a permutation and* $\mathbf{h}(\mathbf{s}) = (h_1(s_1), ..., h_n(s_n))$ *an invertible element-wise function.*

Property *(i)* is the distribution-level analogue to *(i)* of Prop. 4.4 and only allows for orthogonality violations on sets of measure zero w.r.t. $p_{\mathbf{s}}$. This means that C_{IMA} can only be zero if $\mathbf{f}$ is an *orthogonal coordinate transformation* almost everywhere [19, 52, 66], see Fig. 3 for an example. We particularly stress property *(ii)*, as it precisely matches the inherent indeterminacy of nonlinear ICA: C_{IMA} *is blind to reparametrisation of the sources by permutation and element wise transformation.*

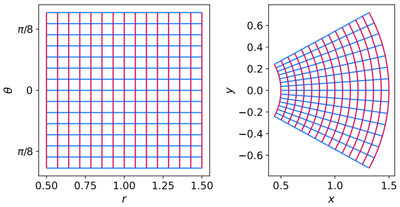


Figure 3: An example of a (non-conformal) orthogonal coordinate transformation from polar (left) to Cartesian (right) coordinates.

4.3 Theoretical analysis and justification of C_{IMA}

We now show that, under suitable assumptions on the generative model (1), a large class of spurious solutions—such as those based on the Darmois construction (4) or measure preserving automorphisms such as $\mathbf{a}^{\mathbf{R}}$ from (5) as described in § 2.1—exhibit nonzero IMA contrast. Denote the class of nonlinear ICA models satisfying (7) (IMA) by $\mathcal{M}_{\text{IMA}} = \{(\mathbf{f}, p_{\mathbf{s}}) \in \mathcal{F} \times \mathcal{P} : C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}}) = 0\} \subset \mathcal{F} \times \mathcal{P}$. Our first main theoretical result is that, under mild assumptions on the observations, Darmois solutions will have strictly positive C_{IMA}, making them distinguishable from those in $\mathcal{M}_{\text{IMA}}$.

Theorem 4.7. *Assume the data generating process in* (1) *and assume that* $x_i \not\perp\!\!\!\perp x_j$ *for some* $i \neq j$*. Then any Darmois solution* $(\mathbf{f}^D, p_{\mathbf{u}})$ *based on* $\mathbf{g}^D$ *as defined in* (4) *satisfies* $C_{\text{IMA}}(\mathbf{f}^D, p_{\mathbf{u}}) > 0$*. Thus a solution satisfying* $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}}) = 0$ *can be distinguished from* $(\mathbf{f}^D, p_{\mathbf{u}})$ *based on the contrast* C_{IMA}*.*

The proof is based on the fact that the Jacobian of $\mathbf{g}^{\mathrm{D}}$ is triangular (see Remark 2.4) and on the specific form of (4). A specific example of a mixing process satisfying the IMA assumption is the case where $\mathbf{f}$ is a conformal (angle-preserving) map.

Definition 4.8 (Conformal map)**.** A smooth map $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ is conformal if $\mathbf{J}_{\mathbf{f}}(\mathbf{s}) = \mathbf{O}(\mathbf{s})\lambda(\mathbf{s})\ \forall \mathbf{s}$, where $\lambda : \mathbb{R}^n \to \mathbb{R}$ is a scalar field, and $\mathbf{O} \in O(n)$ is an orthogonal matrix.

Corollary 4.9. *Under assumptions of Thm. 4.7, if additionally* $\mathbf{f}$ *is a conformal map, then* $(\mathbf{f}, p_{\mathbf{s}}) \in \mathcal{M}_{\text{IMA}}$ *for any* $p_{\mathbf{s}} \in \mathcal{P}$ *due to Prop. 4.6 (i), see Defn. 4.8. Based on Thm. 4.7,* $(\mathbf{f}, p_{\mathbf{s}})$ *is thus distinguishable from Darmois solutions* $(\mathbf{f}^D, p_{\mathbf{u}})$*.*

This is consistent with a result that proves identifiability of conformal maps for $n = 2$ and conjectures it in general [39].[12] However, conformal maps are only a small subset of all maps for which $C_{\text{IMA}} = 0$, as is apparent from the more flexible condition of Prop. 4.6 *(i)*, compared to the stricter Defn. 4.8.

[12] Note that Corollary 4.9 holds for any dimensionality n.

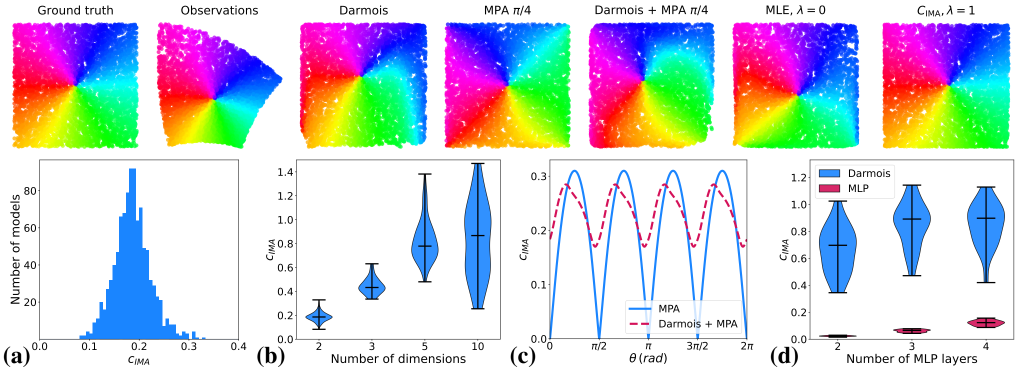


Figure 4: **Top.** Visual comparison of different nonlinear ICA solutions for $n = 2$: *(left to right)* true sources; observed mixtures; Darmois solution; true unmixing, composed with the measure preserving automorphism (MPA) from (5) (with rotation by $^{\pi}/_{4}$); Darmois solution composed with the same MPA; maximum likelihood ($\lambda = 0$); and C_{IMA}-regularised approach ($\lambda = 1$). **Bottom.** Quantitative comparison of C_{IMA} for different spurious solutions: learnt Darmois solutions for **(a)** $n = 2$, and **(b)** $n \in \{2, 3, 5, 10\}$ dimensions; **(c)** composition of the MPA (5) in $n = 2$ dim. with the true solution (blue) and a Darmois solution (red) for different angles. **(d)** C_{IMA} distribution for true MLP mixing (red) vs. Darmois solution (blue) for $n = 5$ dim., $L \in \{2, 3, 4\}$ layers.

Example 4.10 (Polar to Cartesian coordinate transform). Consider the *non-conformal* transformation from polar to Cartesian coordinates (see Fig. 3), defined as $(x, y) = \mathbf{f}(r, \theta) := (r\cos(\theta), r\sin(\theta))$ with independent sources $\mathbf{s} = (r, \theta)$, with $r \sim U(0, R)$ and $\theta \sim U(0, 2\pi)$.[13] Then, $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}}) = 0$ and $C_{\text{IMA}}(\mathbf{f}^{\text{D}}, p_{\mathbf{u}}) > 0$ for any Darmois solution $(\mathbf{f}^{\text{D}}, p_{\mathbf{u}})$ —see Appendix D for details.

Finally, for the case in which the true mixing is linear, we obtain the following result.

Corollary 4.11. *Consider a linear ICA model, $\mathbf{x} = \mathbf{A}\mathbf{s}$, with $\mathbb{E}[\mathbf{s}^{\top}\mathbf{s}] = \mathbf{I}$, and $\mathbf{A} \in O(n)$ an orthogonal, non-trivial mixing matrix, i.e., not the product of a diagonal and a permutation matrix $\mathbf{DP}$. If at most one of the s_i is Gaussian, then $C_{\text{IMA}}(\mathbf{A}, p_{\mathbf{s}}) = 0$ and $C_{\text{IMA}}(\mathbf{f}^{D}, p_{\mathbf{u}}) > 0$.*

In a "blind" setting, we may not know a priori whether the true mixing is linear or not, and thus choose to learn a nonlinear unmixing. Corollary 4.11 shows that, in this case, Darmois solutions are still distinguishable from the true mixing via C_{IMA}. Note that unlike in Corollary 4.9, the assumption that $x_i \not\perp\!\!\!\perp x_j$ for some $i \neq j$ is not required for Corollary 4.11. In fact, due to Theorem 11 of [17], it follows from the assumed linear ICA model with non-Gaussian sources, and the fact that the mixing matrix is not the product of a diagonal and a permutation matrix (see also Appendix A).

Having shown that the IMA principle allows to distinguish a class of models (including, but not limited to conformal maps) from Darmois solutions, we next turn to a second well-known counterexample to identifiability: the "rotated-Gaussian" MPA $\mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}})$ (5) from Defn. 2.5. Our second main theoretical result is that, under suitable assumptions, this class of MPAs can also be ruled out for "non-trivial" $\mathbf{R}$.

Theorem 4.12. *Let $(\mathbf{f}, p_{\mathbf{s}}) \in \mathcal{M}_{\text{IMA}}$ and assume that $\mathbf{f}$ is a conformal map. Given $\mathbf{R} \in O(n)$, assume additionally that $\exists$ at least one non-Gaussian s_i whose associated canonical basis vector $\mathbf{e}_i$ is not transformed by $\mathbf{R}^{-1} = \mathbf{R}^{\top}$ into another canonical basis vector $\mathbf{e}_j$. Then $C_{\text{IMA}}(\mathbf{f} \circ \mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}}), p_{\mathbf{s}}) > 0$.*

Thm. 4.12 states that for conformal maps, applying the $\mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}})$ transformation at the level of the sources leads to an increase in C_{IMA} except for very specific rotations $\mathbf{R}$ that are "fine-tuned" to $p_{\mathbf{s}}$ in the sense that they permute all non-Gaussian sources s_i with another s_j. Interestingly, as for the linear case, non-Gaussianity again plays an important role in the proof of Thm. 4.12.

5 Experiments

Our theoretical results from § 4 suggest that C_{IMA} is a promising contrast function for nonlinear blind source separation. We test this empirically by evaluating the C_{IMA} of spurious nonlinear ICA solutions (§ 5.1), and using it as a learning objective to recover the true solution (§ 5.2).

We sample the ground truth sources from a uniform distribution in $[0, 1]^n$; the reconstructed sources are also mapped to the uniform hypercube as a reference measure via the CDF transform. Unless

[13] For different $p_{\mathbf{s}}$, (x, y) can be made to have independent Gaussian components ([98], II.B), and C_{IMA}-identifiability is lost; this shows that the assumption of Thm. 4.7 that $x_i \not\perp\!\!\!\perp x_j$ for some $i \neq j$ is crucial.

otherwise specified, the ground truth mixing $\mathbf{f}$ is a Möbius transformation [81] (i.e., a conformal map) with randomly sampled parameters, thereby satisfying Principle 4.1. In all of our experiments, we use JAX [12] and Distrax [13]. For additional technical details, equations and plots see Appendix E. The code to reproduce our experiments is available at this link.

5.1 Numerical evaluation of the C_{IMA} contrast for spurious nonlinear ICA solutions

Learning the Darmois construction. To learn the Darmois construction from data, we use normalising flows, see [35, 69]. Since Darmois solutions have triangular Jacobian (Remark 2.4), we use an architecture based on residual flows [16] which we constrain such that the Jacobian of the full model is triangular. This yields an expressive model which we train effectively via maximum likelihood.

C_{IMA} of Darmois solutions. To check whether Darmois solutions (learnt from finite data) can be distinguished from the true one, as predicted by Thm. 4.7, we generate 1000 random mixing functions for $n = 2$, compute the C_{IMA} values of learnt solutions, and find that all values are indeed significantly larger than zero, see Fig. 4 **(a)**. The same holds for higher dimensions, see Fig. 4 **(b)** for results with 50 random mixings for $n \in \{2, 3, 5, 10\}$: with higher dimensionality, both the mean and variance of the C_{IMA} distribution for the learnt Darmois solutions generally attain higher values.[14] We confirmed these findings for mappings which are not conformal, while still satisfying (7), in Appendix E.5.

C_{IMA} of MPAs. We also investigate the effect on C_{IMA} of applying an MPA $\mathbf{a}^{\mathbf{R}}(\cdot)$ from (5) to the true solution or a learnt Darmois solution. Results for $n = 2$ dim. for different rotation matrices $\mathbf{R}$ (parametrised by the angle θ) are shown in Fig. 4 **(c)**. As expected, the behavior is periodic in θ, and vanishes for the true solution (blue) at multiples of $\pi/2$, i.e., when $\mathbf{R}$ is a permutation matrix, as predicted by Thm. 4.12. For the learnt Darmois solution (red, dashed) C_{IMA} remains larger than zero.

C_{IMA} values for random MLPs. Lastly, we study the behavior of spurious solutions based on the Darmois construction under deviations from our assumption of $C_{\text{IMA}} = 0$ for the true mixing function. To this end, we use invertible MLPs with orthogonal weight initalisation and `leaky_tanh` activations [29] as mixing functions; the more layers L are added to the mixing MLP, the larger a deviation from our assumptions is expected. We compare the true mixing and learnt Darmois solutions over 20 realisations for each $L \in \{2, 3, 4\}$, $n = 5$. Results are shown in figure Fig. 4 **(d)**: the C_{IMA} of the mixing MLPs grows with L; still, the one of the Darmois solution is typically higher.

Summary. We verify that spurious solutions can be distinguished from the true one based on C_{IMA}.

5.2 Learning nonlinear ICA solutions with C_{IMA}-regularised maximum likelihood

Experimental setup. To use C_{IMA} as a learning signal, we consider a regularised maximum-likelihood approach, with the following objective: $\mathcal{L}(\mathbf{g}) = \mathbb{E}_{\mathbf{x}}[\log p_{\mathbf{g}}(\mathbf{x})] - \lambda\, C_{\text{IMA}}(\mathbf{g}^{-1}, p_{\mathbf{y}})$, where $\mathbf{g}$ denotes the learnt unmixing, $\mathbf{y} = \mathbf{g}(\mathbf{x})$ the reconstructed sources, and $\lambda \geq 0$ a Lagrange multiplier. For $\lambda = 0$, this corresponds to standard maximum likelihood estimation, whereas for $\lambda > 0$, $\mathcal{L}$ lower-bounds the likelihood, and recovers it exactly iff. $(\mathbf{g}^{-1}, p_{\mathbf{y}}) \in \mathcal{M}_{\text{IMA}}$. We train a residual flow $\mathbf{g}$ (with full Jacobian) to maximise $\mathcal{L}$. For evaluation, we compute (i) the KL divergence to the true data likelihood, as a measure of goodness of fit for the learnt flow model; and (ii) the mean correlation coefficient (MCC) between ground truth and reconstructed sources [37, 49]. We also introduce (iii) a nonlinear extension of the Amari distance [5] between the true mixing and the learnt unmixing, which is larger than or equal to zero, with equality iff. the learnt model belongs to the BSS equivalence class (Defn. 2.2) of the true solution, see Appendix E.5 for details.

Results. In Fig. 4 *(Top)*, we show an example of the distortion induced by different *spurious* solutions for $n = 2$, and contrast it with a solution learnt using our proposed objective *(rightmost plot)*. Visually, we find that the C_{IMA}-regularised solution (with $\lambda = 1$) recovers the true sources most faithfully. Quantitative results for 50 learnt models for each $\lambda \in \{0.0, 0.5, 1.0\}$ and $n \in \{5, 7\}$ are summarised in Fig. 5 (see Appendix E for additional plots) . As indicated by the KL divergence values *(left)*, most trained models achieve a good fit to the data across all values of λ.[15] We observe that using C_{IMA} (i.e., $\lambda > 0$) is beneficial for BSS, both in terms of our nonlinear Amari distance *(center, lower is better)* and MCC *(right, higher is better)*, though we do not observe a substantial difference between $\lambda = 0.5$ and $\lambda = 1$.[16]

Summary: C_{IMA} can be a useful learning signal to recover the true solution.

[14] the latter possibly due to the increased difficulty of the learning task for larger n

[15] models with $n = 7$ have high outlier KL values, seemingly less pronounced for nonzero values of λ

[16] In Appendix E.5, we also show that our method is superior to a linear ICA baseline, FastICA [36].

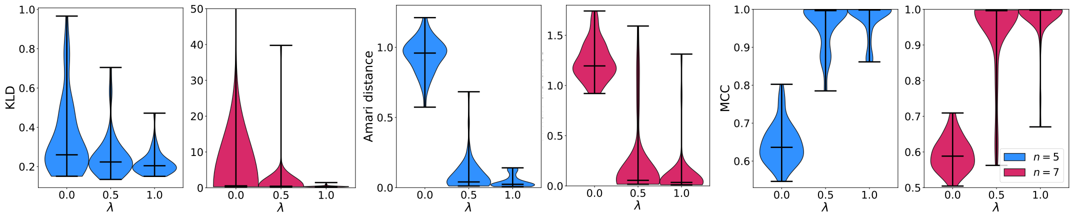


Figure 5: BSS via C_{IMA}-regularised MLE for, side by side, $n = 5$ (blue) and $n = 7$ (red) dim. with $\lambda \in \{0.0, 0.5, 1.0\}$. *(Left)* KL-divergence between ground truth likelihood and learnt model; *(center)* nonlinear Amari distance given true mixing and learnt unmixing; *(right)* MCC between true and reconstructed sources.

6 Discussion

Assumptions on the mixing function. Instead of relying on weak supervision in the form of auxiliary variables [28, 30, 37, 38, 41, 49], our IMA approach places additional constraints on the functional form of the mixing process. In a similar vein, the *minimal nonlinear distortion principle* [108] proposes to favor solutions that are as close to linear as possible. Another example is the *post-nonlinear model* [98, 109], which assumes an element-wise nonlinearity applied after a linear mixing. IMA is different in that it still allows for strongly nonlinear mixings (see, e.g., Fig. 3) provided that the columns of their Jacobians are (close to) orthogonal. In the related field of disentanglement [8, 58], a line of work that focuses on image generation with adversarial networks [24] similarly proposes to constrain the "generator" function via regularisation of its Jacobian [82] or Hessian [74], though mostly from an empirically-driven, rather than from an identifiability perspective as in the present work.

Towards identifiability with C_{IMA}. The IMA principle rules out a large class of spurious solutions to nonlinear ICA. While we do not present a full identifiability result, our experiments show that C_{IMA} can be used to recover the BSS equivalence class, suggesting that identifiability might indeed hold, possibly under additional assumptions—e.g., for conformal maps [39].

IMA and independence of cause and mechanism. While inspired by measures of independence of cause and mechanism as traditionally used for cause-effect inference [18, 45, 46, 110], we view the IMA principle as addressing a different question, in the sense that they evaluate independence between different elements of the causal model. Any nonlinear ICA solution that satisfies the IMA Principle 4.1 can be turned into one with uniform reconstructed sources—thus satisfying IGCI as argued in § 3—through composition with an element-wise transformation which, according to Prop. 4.6 *(ii)*, leaves the C_{IMA} value unchanged. Both IGCI (6) and IMA (7) can therefore be fulfilled simultaneosly, while the former on its own is inconsequential for BSS as shown in Prop. 3.1.

BSS through algorithmic information. Algorithmic information theory has previously been proposed as a unifying framework for identifiable approaches to *linear* BSS [67, 68], in the sense that commonly-used contrast functions could, under suitable assumptions, be interpreted as proxies for the total complexity of the mixing and the reconstructed sources. However, to the best of our knowledge, the problem of specifying suitable proxies for the complexity of *nonlinear* mixing functions has not yet been addressed. We conjecture that our framework could be linked to this view, based on the additional assumption of algorithmic independence of causal mechanisms [43], thus potentially representing an approach to *nonlinear* BSS by minimisation of algorithmic complexity.

ICA for causal inference & causality for ICA. Past advances in ICA have inspired novel causal discovery methods [50, 64, 92]. The present work constitutes, to the best of our knowledge, the first effort to use ideas from causality (specifically ICM) for BSS. An application of the IMA principle to causal discovery or causal representation learning [88] is an interesting direction for future work.

Conclusion. We introduce IMA, a path to nonlinear BSS inspired by concepts from causality. We postulate that the *influences* of different sources on the observed distribution should be approximately independent, and formalise this as an orthogonality condition on the columns of the Jacobian. We prove that this constraint is generally violated by well-known spurious nonlinear ICA solutions, and propose a regularised maximum likelihood approach which we empirically demonstrate to be effective in recovering the true solution. Our IMA principle holds exactly for orthogonal coordinate transformations, and is thus of potential interest for learning spatial representations [33], robot dynamics [63], or physics problems where orthogonal reference frames are common [66].

Acknowledgements

The authors thank Aapo Hyvärinen, Adrián Javaloy Bornás, Dominik Janzing, Giambattista Parascandolo, Giancarlo Fissore, Nasim Rahaman, Patrick Burauel, Patrik Reizinger, Paul Rubenstein, Shubhangi Ghosh, and the anonymous reviewers for helpful comments and discussions.

Funding Transparency Statement

This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B; and by the Machine Learning Cluster of Excellence, EXC number 2064/1 - Project number 390727645.

References

[1] Pierre Ablin, Jean-François Cardoso, and Alexandre Gramfort. Faster independent component analysis by preconditioning with hessian approximations. *IEEE Transactions on Signal Processing*, 66(15): 4040–4049, 2018.

[2] Khaled Alyani, Marco Congedo, and Maher Moakher. Diagonality measures of Hermitian positive-definite matrices with application to the approximate joint diagonalization problem. *Linear Algebra and its Applications*, 528:290–320, 2017.

[3] Shun-ichi Amari. Information geometry. *Japanese Journal of Mathematics*, 16(1):1–48, 2021.

[4] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

[5] Shun-ichi Amari, Andrzej Cichocki, Howard Hua Yang, et al. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems*, pages 757–763, 1996.

[6] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[7] Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems*, 28:1342–1350, 2015.

[8] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[9] M Besserve, N Shajarisales, B Schölkopf, and D Janzing. Group invariance principles for causal generative models. In *21st International Conference on Artificial Intelligence and Statistics (AISTATS 2018)*, pages 557–565. International Machine Learning Society, 2018.

[10] M Besserve, A Mehrjou, R Sun, and B Schölkopf. Counterfactuals uncover the modular structure of deep generative models. In *Eighth International Conference on Learning Representations (ICLR 2020)*, 2020.

[11] Patrick Blöbaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schölkopf. Cause-effect inference by comparing regression errors. In *International Conference on Artificial Intelligence and Statistics*, pages 900–909. PMLR, 2018.

[12] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, and Skye Wanderman-Milne. JAX: composable transformations of Python+NumPy programs, 2018.

[13] Jake Bruce, David Budden, Matteo Hessel, George Papamakarios, and Francisco Ruiz. Distrax: Probability distributions in JAX, 2021.

[14] Lars Buesing, Theophane Weber, Yori Zwols, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. *arXiv preprint arXiv:1811.06272*, 2018.

[15] Jean-François Cardoso. The three easy routes to independent component analysis; contrasts and geometry. In *Proc. ICA*, volume 2001, 2001.

[16] Ricky T. Q. Chen, Jens Behrmann, David Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. In *Advances in Neural Information Processing Systems*, 2019.

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

[18] Povilas Daniušis, Dominik Janzing, Joris Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang, and Bernhard Schölkopf. Inferring deterministic causal relations. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 143–150, 2010.

[19] Gaston Darboux. *Leçons sur les systemes orthogonaux et les coordonnées curvilignes*. Gauthier-Villars, 1910.

[20] George Darmois. Analyse des liaisons de probabilité. In *Proc. Int. Stat. Conferences 1947*, page 231, 1951.

[21] George Darmois. Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut international de statistique*, pages 2–8, 1953.

[22] Andrew Forney, Judea Pearl, and Elias Bareinboim. Counterfactual data-fusion for online reinforcement learners. In *International Conference on Machine Learning*, pages 1156–1164. PMLR, 2017.

[23] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International Conference on Machine Learning*, pages 2839–2848. PMLR, 2016.

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[25] Alexander N Gorban and Ivan Yu Tyukin. Blessing of dimensionality: mathematical foundations of the statistical physics of data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2118):20170237, 2018.

[26] Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. Recurrent independent mechanisms. In *9th International Conference on Learning Representations*, 2021.

[27] Daniel Greenfeld and Uri Shalit. Robust learning with the hilbert-schmidt independence criterion. In *International Conference on Machine Learning*, pages 3759–3768. PMLR, 2020.

[28] Luigi Gresele, Paul K Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The Incomplete Rosetta Stone problem: Identifiability results for multi-view nonlinear ICA. In *Uncertainty in Artificial Intelligence*, pages 217–227. PMLR, 2019.

[29] Luigi Gresele, Giancarlo Fissore, Adrián Javaloy, Bernhard Schölkopf, and Aapo Hyvarinen. Relative gradient optimization of the Jacobian term in unsupervised deep learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[30] Hermanni Hälvä and Aapo Hyvärinen. Hidden Markov nonlinear ICA: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pages 939–948. PMLR, 2020.

[31] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2):303–348, 2021.

[32] Tom Hennigan, Trevor Cai, Tamara Norman, and Igor Babuschkin. Haiku: Sonnet for JAX, 2020.

[33] Geoffrey E Hinton and Lawrence M Parsons. Frames of reference and mental imagery. *Attention and performance IX*, pages 261–277, 1981.

[34] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems*, 21:689–696, 2008.

[35] Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron C. Courville. Neural autoregressive flows. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 2083–2092. PMLR, 2018.

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

[37] Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, pages 3765–3773, 2016.

[38] Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

[39] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

[40] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, Ltd, 2001.

[41] Aapo Hyvärinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

[42] Dominik Janzing. Causal version of principle of insufficient reason and maxent. *arXiv preprint arXiv:2102.03906*, 2021.

[43] Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

[44] Dominik Janzing and Bernhard Schölkopf. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1), 2018.

[45] Dominik Janzing, Patrik O Hoyer, and Bernhard Schölkopf. Telling cause from effect based on high-dimensional observations. In *International Conference on Machine Learning*, 2010.

[46] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

[47] Dominik Janzing, Bastian Steudel, Naji Shajarisales, and Bernhard Schölkopf. Justifying information-geometric causal inference. In *Measures of complexity*, pages 253–265. Springer, 2015.

[48] Dominik Janzing, Raphael Chaves, and Bernhard Schölkopf. Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference. *New Journal of Physics*, 18(9): 093052, 2016.

[49] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[50] Ilyes Khemakhem, Ricardo Monti, Robert Leech, and Aapo Hyvarinen. Causal autoregressive flows. In *International Conference on Artificial Intelligence and Statistics*, pages 3520–3528. PMLR, 2021.

[51] Andrei N Kolmogorov. On tables of random numbers. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 369–376, 1963.

[52] Gabriel Lamé. *Leçons sur les coordonnées curvilignes et leurs diverses applications*. Mallet-Bachelier, 1859.

[53] Sanghack Lee and Elias Bareinboim. Structural causal bandits: where to intervene? *Advances in Neural Information Processing Systems 31*, 31, 2018.

[54] Felix Leeb, Yashas Annadani, Stefan Bauer, and Bernhard Schölkopf. Structural autoencoders improve representations for generation and transfer. *arXiv preprint arXiv:2006.07796*, 2020.

[55] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[56] Jan Lemeire and Erik Dirkx. Causal models as minimal descriptions of multivariate systems, 2006.

[57] Jan Lemeire and Dominik Janzing. Replacing causal faithfulness with algorithmic independence of conditionals. *Minds and Machines*, 23(2):227–249, 2013.

[58] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124. PMLR, 2019.

[59] Chaochao Lu, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Deconfounding reinforcement learning in observational settings. *arXiv preprint arXiv:1812.10576*, 2018.

[60] Chaochao Lu, Biwei Huang, Ke Wang, José Miguel Hernández-Lobato, Kun Zhang, and Bernhard Schölkopf. Sample-efficient reinforcement learning via counterfactual-based data augmentation. *arXiv preprint arXiv:2012.09092*, 2020.

[61] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 10869–10879, 2018.

[62] James Mahoney. Beyond correlational analysis: Recent innovations in theory and method. In *Sociological forum*, pages 575–593. JSTOR, 2001.

[63] Michael Mistry, Jonas Buchli, and Stefan Schaal. Inverse dynamics control of floating base systems using orthogonal decomposition. In *2010 IEEE International Conference on Robotics and Automation*, pages 3406–3412. IEEE, 2010.

[64] Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. Causal discovery with general non-linear relationships using non-linear ICA. In *Uncertainty in Artificial Intelligence*, pages 186–195. PMLR, 2020.

[65] Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

[66] Parry Moon and Domina Eberle Spencer. *Field theory handbook, including coordinate systems, differential equations and their solutions*. Springer, 1971.

[67] Petteri Pajunen. Blind source separation using algorithmic information theory. *Neurocomputing*, 22(1-3): 35–48, 1998.

[68] Petteri Pajunen. Blind source separation of natural signals based on approximate complexity minimization. In *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, page 267. Citeseer, 1999.

[69] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[70] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *International Conference on Machine Learning*, pages 4036–4044. PMLR, 2018.

[71] Judea Pearl. *Causality*. Cambridge university press, 2009.

[72] Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014.

[73] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[74] William S. Peebles, John Peebles, Jun-Yan Zhu, Alexei A. Efros, and Antonio Torralba. The Hessian penalty: A weak prior for unsupervised disentanglement. In *ECCV*, volume 12351, pages 581–597. Springer, 2020.

[75] Jonas Peters and Peter Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

[76] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

[77] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

[78] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[79] Dinh-Tuan Pham and J-F Cardoso. Blind separation of instantaneous mixtures of nonstationary sources. *IEEE Transactions on Signal Processing*, 49(9):1837–1848, 2001.

[80] Dinh Tuan Pham and Philippe Garat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Transactions on Signal Processing*, 45(7):1712–1725, 1997.

[81] Robert Phillips. Liouville's theorem. *Pacific Journal of Mathematics*, 28(2):397–405, 1969.

[82] Aditya Ramesh, Youngduck Choi, and Yann LeCun. A spectral regularizer for unsupervised disentanglement. *arXiv preprint arXiv:1812.01161*, 2018.

[83] Danilo Jimenez Rezende. Short notes on divergence measures, 2018.

[84] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

[85] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021.

[86] Wesley C Salmon. *Scientific explanation and the causal structure of the world*. Princeton University Press, 2020.

[87] B Schölkopf, D Janzing, J Peters, E Sgouritsa, K Zhang, and J Mooij. On causal and anticausal learning. In *29th International Conference on Machine Learning (ICML 2012)*, pages 1255–1262. International Machine Learning Society, 2012.

[88] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021.

[89] Manfred R Schroeder. Listening with two ears. *Music perception*, 10(3):255–280, 1993.

[90] Naji Shajarisales, Dominik Janzing, Bernhard Schoelkopf, and Michel Besserve. Telling cause from effect in deterministic linear dynamical systems. In *International Conference on Machine Learning*, pages 285–294. PMLR, 2015.

[91] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Disentangled generative causal representation learning. *arXiv preprint arXiv:2010.02637*, 2020.

[92] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[93] VP Skitović. On a property of a normal distribution. In *Doklady Akad. Nauk*, 1953.

[94] James R Smart. *Modern geometries*. Brooks/Cole Pacific Grove, CA, 1998.

[95] Mirjam Soeten. *Conformal maps and the theorem of Liouville*. PhD thesis, Faculty of Science and Engineering, 2011.

[96] B. Steudel, D. Janzing, and B. Schölkopf. Causal Markov condition for submodular information measures. In A. Kalai and M. Mohri, editors, *Conference on Learning Theory (COLT)*, pages 464–476, Madison, WI, USA, 2010. OmniPress.

[97] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019.

[98] Anisse Taleb and Christian Jutten. Source separation in post-nonlinear mixtures. *IEEE Transactions on Signal Processing*, 47(10):2807–2820, 1999.

[99] Charles Tilly. Historical analysis of political processes. In *Handbook of sociological theory*, pages 567–588. Springer, 2001.

[100] Ruy Tojeiro. Liouville's theorem revisited. *Enseignement Mathematique*, 53(1/2):67, 2007.

[101] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[102] Julius von Kügelgen, Alexander Mey, and Marco Loog. Semi-generative modelling: Covariate-shift adaptation with cause and effect features. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1361–1369. PMLR, 2019.

[103] Julius von Kügelgen, Alexander Mey, Marco Loog, and Bernhard Schölkopf. Semi-supervised learning, causality, and the conditional cluster assumption. In *Conference on Uncertainty in Artificial Intelligence*, pages 1–10. PMLR, 2020.

[104] Julius von Kügelgen, Ivan Ustyuzhaninov, Peter Gehler, Matthias Bethge, and Bernhard Schölkopf. Towards causal generative scene models via competition of experts. In *ICLR 2020 Workshop on "Causal Learning for Decision Making"*, 2020.

[105] Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, 2021.

[106] Junzhe Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. *Advances in Neural Information Processing Systems*, 32:13401–13411, 2019.

[107] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 819–827, 2013.

[108] Kun Zhang and Laiwan Chan. Minimal nonlinear distortion principle for nonlinear independent component analysis. *Journal of Machine Learning Research*, 9(Nov):2455–2487, 2008.

[109] Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 647–655. AUAI Press, 2009.

[110] Jakob Zscheischler, Dominik Janzing, and Kun Zhang. Testing whether linear equations are causal: A free probability theory approach. In *27th Conference on Uncertainty in Artificial Intelligence*, pages 839–847, 2011.

Checklist

1. For all authors...

   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

   (b) Did you describe the limitations of your work? [Yes] See § 6, where we discuss limitations of our theory (e.g. lines 354-357) and open questions.

   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work is mainly theoretical, and we believe it does not bear immediate negative societal impacts.

   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We formally define the problem setting in § 2 and § 3, and transparently state and discuss our assumptions in § 4.

   (b) Did you include complete proofs of all theoretical results? [Yes] Due to space constraints, full proofs and detailed explanations are mainly reported in appendix C and appendix D; the proof of Prop. 3.1 is given in the main text.

3. If you ran experiments...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code, data, and the configuration files are included in the supplemental material.

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All details, including hyperparameters, seed for random number generators, etc. are specified in the configuration files which will be included in the supplemental.

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We visualized the distribution of the considered quantities via histograms and violin plots, see e.g. Figure 4 and Figure 5.

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] They are specified appendix E.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes] We use the Python libraries JAX and Distrax and cited the creators in the article.

   (b) Did you mention the license of the assets? [Yes] Both packages have Apache License 2.0; we report this in the appendices.

   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Implementations of our proposed methods and metrics will be provided in the supplemental.

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]