Independent mechanism analysis, a new concept?

Causality Reading Group

Luigi Gresele

February 4, 2023

MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS

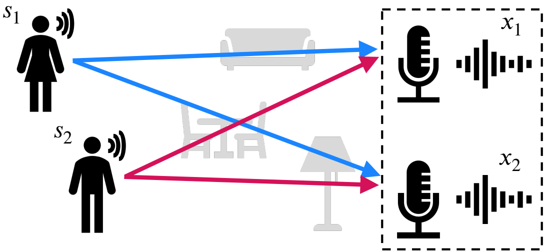
Overview

- **Part I:** An introduction to identifiable representation learning.
- **Part II:** A new principle for identifiable representation learning, inspired by the principle of *independent causal mechanisms*.
  (i) *Independent mechanism analysis, a new concept?*
  (ii) *Embrace the Gap: VAEs Perform Independent Mechanism Analysis*

Motivation: Two Metaphors

The Cocktail-Party Problem

Multiple speakers in a room; microphones record mixtures $\mathbf{x} = \mathbf{f}(\mathbf{s})$ of the speakers' voices s_1, s_2 (source signals).

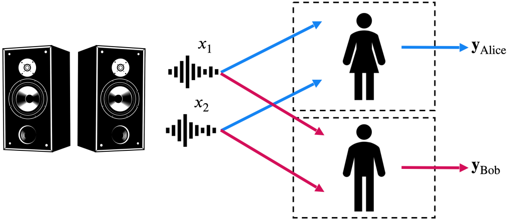


Blind source separation (BSS): Given only the recorded mixtures x, can we recover (separate) the original source signals s_1, s_2?

The Independent-Listeners Problem

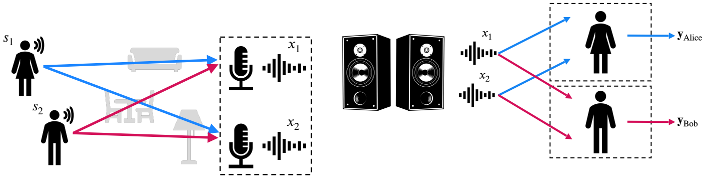
Two listeners, Alice and Bob, are exposed to the same audio signals $\mathbf{x}$.

Each produces their own attempt at blind source separation—$\mathbf{y}_{\text{Alice}} = \mathbf{g}_{\text{Alice}}(\mathbf{x})$ for Alice and $\mathbf{y}_{\text{Bob}} = \mathbf{g}_{\text{Bob}}(\mathbf{x})$ for Bob.



Are the two vectors $\mathbf{y}_{\text{Alice}}$ and $\mathbf{y}_{\text{Bob}}$ equal? If not, how are they related?

Difference Between the Two Metaphors



- Studying the similarity of representations learned by different intelligent systems might be relevant beyond the question whether such representations faithfully reconstruct some ground-truth and invert the true data-generating process—which is the objective of blind source separation.
- Alice and Bob might both *fail* the source separation task; it might even be unclear whether there is a *correct* answer.
- What we are interested in is whether they will provide the *same* answer (up to some specified ambiguities).

Motivation: Learning Representations in Machines and the Brain

What is a Representation? i

Starting from the definition by Marr (2010):

> *A* representation *is a formal system for making explicit certain entities or types of information, together with a specification of how the system does it. [...]*

How to interpret this in the light of the two metaphors?

- We can think of the vectors $\mathbf{y}_{\text{Alice}} = \mathbf{g}_{\text{Alice}}(\mathbf{x})$ and $\mathbf{y}_{\text{Bob}} = \mathbf{g}_{\text{Bob}}(\mathbf{x})$ as *representations* of the observations $\mathbf{x}$.
- The vector $\mathbf{x}$ may contain the same information as $\mathbf{s}$, but in $\mathbf{s}$ the separation between voices of the two speakers is made explicit.
- Specifying *"how the system does it"*: Specifying (how the system learns) a function $\mathbf{g}$ s.t. $\mathbf{y} = \mathbf{g}(\mathbf{x})$ makes the information we are interested in explicit (i.e., it separates the sources).

What is a Representation? ii

Marr (2010) additionally argues that alternative representations may differ in what information they make explicit:

> *The Arabic, Roman, and binary numeral systems are all formal systems for representing numbers. […] What [the Arabic numeral system] makes explicit is the number's decomposition into powers of* 10*. The binary numeral system's description of the number thirty-seven is* 100101*, and this description makes explicit the number's decomposition into powers of* 2*.*

For example, the decomposition of a number into powers of 2 (respectively 10) can be discussed based on either representation, but it is made explicit in the binary (respectively Arabic) numeral system.

Representations and “Downstream Tasks” (Marr, 2010)

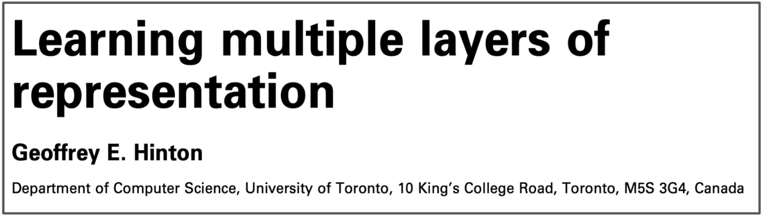
How information is represented can greatly affect how easily we can do things with it:

> *[...] how information is represented can greatly affect how easily it is to do different things with it. This is evident even from our numbers example:* ***It is easy to add, to subtract, and even to multiply if the Arabic or binary representations are used, but it is not at all easy to do these things–especially multiplication–with Roman numerals.*** *This is a key reason why the Roman culture failed to develop mathematics in the way the earlier Arabic cultures had.*

Why is it even Necessary to Talk about Representations?

- If we are interested in a downstream task, should we not simply focus on developing algorithms which are good at solving the task we are interested in (Huszár, 2018)?
    - Violates Vapnik's principle *"never to solve a problem which is more general than the one we actually need to solve"* (Vapnik, 2013).
- The downstream task may be a priori unclear: cocktail party attendees may be unable to predict exactly what questions they will have to ask or answer (possibly part of what makes social interactions interesting).
- For an intelligent system in a complex environment (be it an artificial neural network or a biological party attendee) a good strategy could be to learn representations that *"discard as little information about the data as is practical"* while still *"disentangl[ing] as many factors [of variation] as possible"* (Bengio et al., 2013).

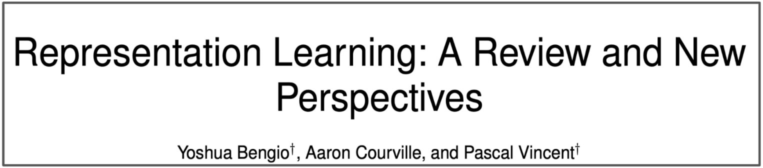
Learning and Representations



Learning multiple layers of representation

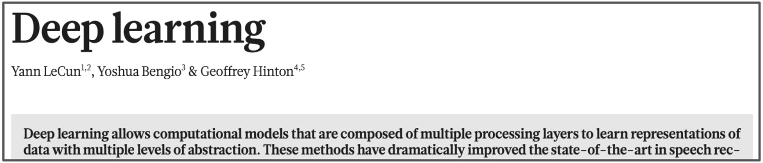
Geoffrey E. Hinton

Department of Computer Science, University of Toronto, 10 King’s College Road, Toronto, M5S 3G4, Canada



Representation Learning: A Review and New Perspectives

Yoshua Bengio†, Aaron Courville, and Pascal Vincent†



Deep learning

Yann LeCun[1,2], Yoshua Bengio[3] & Geoffrey Hinton[4,5]

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech rec-

Learning and Representations

Deep neural networks learn representations which enable solving downstream tasks of interest efficently.

- Hinton (2007):

  > *To achieve its impressive performance in tasks such as speech perception or object recognition, the brain extracts multiple levels of representation from the sensory input.*

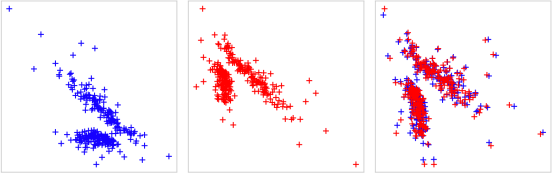
- Bengio et al. (2013):

  > *The success of machine learning algorithms generally depends on data representation, and we hypothesize that this is because different representations can entangle and hide more or less the different explanatory factors of variation behind the data.*

- LeCun et al. (2015):

  > *Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction.*

Example: Independent-Listeners Problem in Deep Learning i



From (Roeder et al., 2021): For a class of objective functions (e.g., deep supervised classification), different representation functions learned on the same data distribution live within linear transformations of each other.

- Comparing representations extracted by neural networks trained “independently” on the same dataset is the objective of theoretical work, e.g., (Khemakhem et al., 2020b; Roeder et al., 2021).

Example: Independent-Listeners Problem in Deep Learning ii

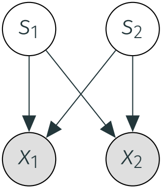
- The question of how representations extracted by different neural network models relate to each other has also been the subject of extensive empirical investigation, see, e.g., (Wu et al., 2020; Moschella et al., 2022).
- The question introduced through the independent-listeners problem may be considered separately from that of identifying a ground truth (cocktail-party): we will introduce a notion (**identifiability**) relevant to investigate both.
- Next, we will introduce a way to formalise and study the cocktail-party problem: **independent component analysis**.

Independent Component Analysis

Independent Components: Estimation vs. Identification i

$$\mathbf{x} = \mathbf{f}(\mathbf{s}), \qquad \text{with smooth, invertible } \mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$$

$$\mathbf{s} \overset{\text{i.i.d.}}{\sim} p_{\mathbf{s}}, \qquad p_{\mathbf{s}}(\mathbf{s}) = \prod_{i=1}^{n} p_{s_i}$$



Independent component estimation:
Given $\mathbf{x}$, find $\mathbf{y} = \mathbf{g}(\mathbf{x})$ s.t. $p(\mathbf{y}) = \prod_{i=1}^{n} p_i(y_i)$.

Blind source separation (BSS):
Given $\mathbf{x}$, reconstruct $\mathbf{s}$ “up to tolerable ambiguities”.

Whether estimating independent components corresponds to solving BSS is related to the concept of *identifiability* of the model class $(\mathbf{f}, p_{\mathbf{s}})$.

Independent Components: Estimation vs. Identification ii

Intuitively, identifiability is the desirable property that *all models which give rise to the same mixture distribution should be "equivalent" up to certain ambiguities*.

Independent component estimation and BSS: Is the model $(\mathbf{f}, p(\mathbf{s}))$ identifiable? Yes if $\mathbf{y}$ is an estimate of $\mathbf{s}$ "up to tolerable ambiguities",

e.g., if $\mathbf{f}$ linear (an invertible matrix $\mathbf{A}$):

- permutation of the sources;
- scale (and sign) of the sources.

Note: As long as information from different sources is not mixed in the reconstructed ones, we are happy with the ambiguities above.

Linear Independent Component Analysis: Identification

Linear ICA as a Case Study in Identifiability

We refer the reader to, e.g., Hyvärinen et al. (2001); Hyvärinen and Oja (2000); Ablin (2019) for comprehensive introductions.

Our main aim is to present ICA as a *case study in identifiability for latent variable models*:

Based on linear ICA, we illustrate the main features of our approach to the study of latent variable models, in particular w.r.t. identifiability.

In the following, we:

(i) introduce the model and discuss what ambiguities should a priori be considered as tolerable;

(ii) characterise its identifiability, including corner cases where it is not achievable.

Linear ICA: Model & Tolerable Ambiguities

$$\mathsf{x} = \mathsf{A}(\mathsf{s}), \qquad \mathsf{A} \in \mathbb{R}^{n \times n}$$

$$\mathsf{s} \overset{\text{i.i.d.}}{\sim} p_{\mathsf{s}}, \qquad p_{\mathsf{s}}(\mathsf{s}) = \prod_{i=1}^{n} p_{s_i}(s_i)\,.$$

Two ambiguities will necessarily hold:

(i) **The variance (and sign) of the latent components cannot be determined.** We have

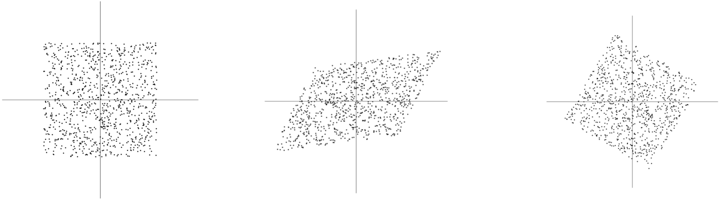
$$\mathsf{x} = \sum_i \left(\frac{1}{\alpha_i}\mathsf{a}_i\right)(s_i \alpha_i)\ ,$$

where α_i is a scalar and a_i denotes the i-th column of A. Customary to fix the source variances, e.g., $\mathbb{E}[\mathsf{s}\mathsf{s}^\top] = \mathsf{I}$.

(ii) **The ordering of the sources cannot be determined.** Given a permutation matrix P, we have $\mathsf{x} = \mathsf{A}\mathsf{s} = \mathsf{A}\mathsf{P}^{-1}\mathsf{P}\mathsf{s}$, with $\mathsf{A}\mathsf{P}^{-1}$ (resp. $\mathsf{P}\mathsf{s}$) a new unmixing matrix (resp. a new set of sources).

Linear ICA Essentially Amounts to Resolving a Rotation

- It can be shown that without loss of generality one can assume that the mixing matrix is orthogonal, $\mathbf{A}\mathbf{A}^\top = \mathbf{I}$.
  - If this does not hold, we can always *whiten* $\mathbf{x}$ first through an invertible linear transformation and obtain an orthogonal mixing (Hyvärinen et al., 2001, §7.4.2).
- The problem of ICA becomes then essentially the problem of estimating an orthogonal matrix—i.e., resolving a rotation.



Left: Uniform, independent sources $\mathbf{s}$; **Center**: Linear mixtures, i.e., observations $\mathbf{x}$; **Right**: Whitened (decorrelated) mixtures.

Identifiability of Linear ICA

We will assume w.l.o.g. that $\mathbf{A}$ is an orthogonal matrix. Suppose that we are given an orthogonal matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$ s.t. the vector

$$\mathbf{y} = \mathbf{B}\mathbf{x} = \mathbf{B}\mathbf{A}\mathbf{s} \tag{1}$$

has independent components. Then $\mathbf{C} = \mathbf{B}\mathbf{A}$ is also orthogonal and the following holds (Darmois, 1953; Skitović, 1953; Comon, 1994).

Theorem (Based on Thm. 11 of Comon (1994))

Let $\mathbf{s}$ *be a vector of n independent components, of which at most one is Gaussian and whose densities are not reduced to a point mass. Let* $\mathbf{C} \in \mathbb{R}^{n \times n}$ *be an orthogonal matrix. Then* $\mathbf{y} = \mathbf{C}\mathbf{s}$ *has (mutually) independent components iff.* $\mathbf{C} = \mathbf{D}\mathbf{P}$, *with* $\mathbf{D}$ *a diagonal matrix and* $\mathbf{P}$ *a permutation matrix.*

ICA solves blind source separation: Linearly transforming $\mathbf{x}$ into independent components is equivalent to separating the sources.

Take-Home Messages from our Case Study

- Certain ambiguities are unavoidable from the outset and deemed “tolerable” (**linear ICA:** permutation and scale).
- An identifiability result defines under which conditions the only unresolvable ambiguities are the tolerable ones.
- This usually holds *apart from a (hopefully small) number of corner cases* which can be characterised (**linear ICA:** more than two Gaussian components).
- Certain assumptions can be made w.l.o.g. given the assumed model (**linear ICA:** orthogonality of the mixing matrix).
    - This might simplify both the theoretical analysis and the estimation (e.g., fewer parameters to estimate).
- Estimation comes with its own problems (e.g., statistical, computational efficiency), which can be addressed separately from the problem of identification.

Identifiability: Formal Definitions

Identifiability in Statistics (Wasserman, 2004; Lehmann and Casella, 2006)

- Identifiability for a class of models parametrised by $\theta \in \Theta$ for observed data $\mathbf{x}$ is the condition that there is a one-to-one mapping between the space of models and the space of parameters.
- The model class is said to be identifiable if

$$\forall \theta, \theta' \in \Theta : \qquad p_\theta(\mathbf{x}) = p_{\theta'}(\mathbf{x}) \quad \forall \mathbf{x} \quad \implies \quad \theta = \theta'. \qquad (2)$$

- The equality on the RHS of (2) is a very strong condition: e.g., for linear ICA, ordering and scale of the sources cannot be determined, so identifiability in the sense of (2) is unattainable.
- **This type of identifiability is impractical for many settings.**

Identifiability in Terms of Equivalence Relations i

The equality on RHS of the implication in (2) can be replaced by an equivalence relation $\sim$ (Khemakhem et al., 2020a).

Definition

An equivalence relation $\sim$ on a set A is a binary relation which satisfies the following three properties:

1. **Reflexivity:** $a \sim a, \forall a \in A$.
2. **Symmetry:** $a \sim b \implies b \sim a, \forall a, b \in A$.
3. **Transitivity:** $(a \sim b) \wedge (b \sim c) \implies a \sim c$.

An equivalence relation on a set A imposes a partition into disjoint subsets, each corresponding to an equivalence class: i.e., the collection of all elements which are $\sim$-related to each other.

Identifiability in Terms of Equivalence Relations ii

For example, $[a] = \{b \in A : a \sim b\}$ denotes the equivalence class containing the element a.

Trivial Example: equality ($=$). For ICA, equivalence up to permutation and scale of the columns of the mixing matrix (see later).

Given an equivalence relation, we can then define the following notion of identifiability: The model class is $\sim$-identifiable if

$$\forall \theta, \theta' \in \Theta : \qquad p_\theta(\mathrm{x}) = p_{\theta'}(\mathrm{x}) \quad \forall \mathrm{x} \qquad \implies \qquad \theta \sim \theta' \,. \tag{3}$$

Defining an appropriate equivalence class for the problem at hand allows us to specify exactly the type of indeterminancies which cannot be resolved and up to which the true generative process can be recovered.

Separate Constraints on the Mixing & Latent Variables i

$$\mathbf{x} = \mathbf{f}(\mathbf{s})\,, \qquad \mathbf{s} \overset{\text{i.i.d.}}{\sim} p_{\mathbf{s}}, \qquad p_{\mathbf{s}}(\mathbf{s}) = \textstyle\prod_{i=1}^{n} p_{s_i}(s_i)\,, \tag{4}$$

Definition ($\sim$-identifiability (Gresele et al., 2021a))

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 (4) is said to be $\sim$*-identifiable on* $\mathcal{M}$ if

$$\forall (\mathbf{f}, p_{\mathbf{s}}), (\tilde{\mathbf{f}}, p_{\tilde{\mathbf{s}}}) \in \mathcal{M} : \qquad \mathbf{f}_* p_{\mathbf{s}} = \tilde{\mathbf{f}}_* p_{\tilde{\mathbf{s}}} \quad \Longrightarrow \quad (\mathbf{f}, p_{\mathbf{s}}) \sim (\tilde{\mathbf{f}}, p_{\tilde{\mathbf{s}}})\,. \tag{5}$$

Separate Constraints on the Mixing & Latent Variables ii

- We use $\mathsf{f} \in \mathcal{F}$ to refer to functions, not parameters. For neural nets, f is the input-output mapping—not, e.g., weights and biases.
- $\mathsf{f}_* p_{\mathsf{s}} = \tilde{\mathsf{f}}_* p_{\tilde{\mathsf{s}}}$ requires that the two models give rise to the same observational distribution.
- For any modification $\mathsf{f} \to \tilde{\mathsf{f}}$, the distribution p_{s} needs to be modified accordingly $p_{\mathsf{s}} \to p_{\tilde{\mathsf{s}}}$ s.t. $\mathsf{f}_* p_{\mathsf{s}}$ and $\tilde{\mathsf{f}}_* p_{\tilde{\mathsf{s}}}$ are the same.
- Separate constraints on the space of mixing functions $\mathcal{F}$ and source distributions $\mathcal{P}$ are expressed more naturally.
    - The former may, e.g., encode that the considered models are linear;
    - the latter specifies the class of latent distributions (e.g., independent components) and exceptions (e.g., more than two Gaussian components);
    - The equivalence class "$\sim$" specifies then "up to" what class of unresolvable ambiguities we can achieve identification.

Example: Identifiability of Linear ICA in this Notation

- $\mathcal{F}_{\text{LIN}}$ is the space of invertible $n \times n$ matrices.
- $\mathcal{P}_{\text{LIN}}$ as the space of source distributions $p_{\mathbf{s}} = \prod_i p_{s_i}$ with at most one Gaussian p_{s_i}.
- Considered models for linear ICA: $\mathcal{M}_{\text{LIN}} = \mathcal{F}_{\text{LIN}} \times \mathcal{P}_{\text{LIN}}$.

Definition ($\sim_{\text{LIN}}$)

The equivalence relation $\sim_{\text{LIN}}$ on $\mathcal{F}_{\text{LIN}} \times \mathcal{P}_{\text{LIN}}$ is given by

$$(\mathbf{A}, p_{\mathbf{s}}) \sim_{\text{LIN}} (\tilde{\mathbf{A}}, p_{\tilde{\mathbf{s}}}) \iff \exists \mathbf{D}, \mathbf{P} \quad \text{s.t.} \quad (\mathbf{A}, p_{\mathbf{s}}) = (\tilde{\mathbf{A}}\mathbf{D}\mathbf{P}, [\mathbf{P}^{-1}\mathbf{D}^{-1}]_* p_{\tilde{\mathbf{s}}}) \,,$$

with $\mathbf{D}$ a diagonal matrix and $\mathbf{P}$ a permutation matrix.

Linear ICA is identifiable up to $\sim_{\text{LIN}}$ on the subspace $\mathcal{M}_{\text{LIN}}$ of pairs of invertible matrices (constraint on $\mathcal{F}$) and factorizing densities for which at most one s_i is Gaussian (constraint on $\mathcal{P}$).

(Non)Identifiability of Nonlinear ICA

Nonlinear ICA Model

The nonlinear ICA model is given by

$$\mathbf{x} = \mathbf{f}(\mathbf{s})\,, \qquad \mathbf{s} \overset{\text{i.i.d.}}{\sim} p_{\mathbf{s}}, \qquad p_{\mathbf{s}}(\mathbf{s}) = \textstyle\prod_{i=1}^{n} p_{s_i}(s_i)\,, \tag{6}$$

with invertible $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ (i.e., no constraints on $\mathcal{F}$ beyond smoothness and invertibility).

Independent component estimation: We want to learn an function $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^n$ s.t. $\mathbf{y} = \mathbf{g}(\mathbf{x})$ has independent components.

Tolerable Ambiguities in Nonlinear ICA i

- The problem is way more severe than in the linear case! Not sufficient to require nongaussianity.
- ~~Tolerable~~ **Unavoidable ambiguities:**
  - **permutation:** as in the linear case;
  - **element-wise nonlinear transformation:** 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, h_j.

Definition ($\sim_{\text{BSS}}$)

The equivalence relation $\sim_{\text{BSS}}$ on $\mathcal{F} \times \mathcal{P}$ is given by

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

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.

Tolerable Ambiguities in Nonlinear ICA ii

- While the ambiguity defined above is larger than that in linear ICA, it is still *tolerable* in the context of blind source separation.
- If we were able to achieve it, we would still *separate* the sources: while they may be nonlinearly distorted w.r.t. the true ones, they would not be mixed in our representation.
- **Unfortunately, for i.i.d. data, we cannot even guarantee reconstruction of the sources up to these ambiguities.**
- This can be shown through suitably constructed counterexamples or "spurious solutions".

The Darmois construction (Darmois, 1953; Hyvärinen and Pajunen, 1999)

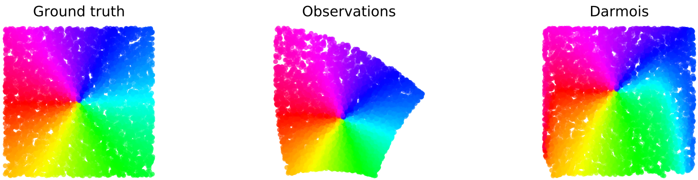
There always $\exists\, \mathbf{g}^{\mathrm{D}} : \mathbb{R}^n \to \mathbb{R}^n$ s.t. $\mathbf{y} = \mathbf{g}(\mathbf{x})$, $p_{\mathbf{y}}(\mathbf{y}) = \prod_{i=1}^{n} p_{y_i}(y_i)$.

Definition (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:

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

The resulting transformation has triangular Jacobian and does not separate the sources, leading to “spurious” solutions.

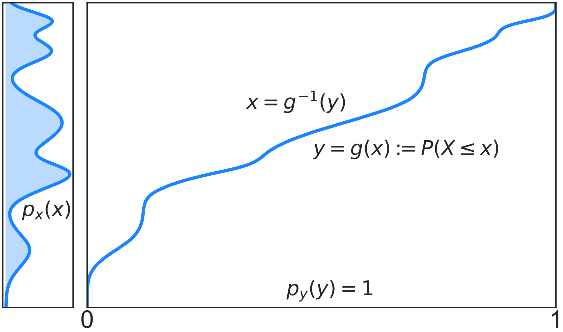


Ground truth

Observations

Darmois

Univariate CDF transform



- CDF: Cumulative density function
- Provides a way to transform any density into the Uniform density on $(0, 1)$.
- The Darmois construction extends this to multivariate, non-factorised densities.

Gaussian-rotated measure preserving automorphism (MPA)

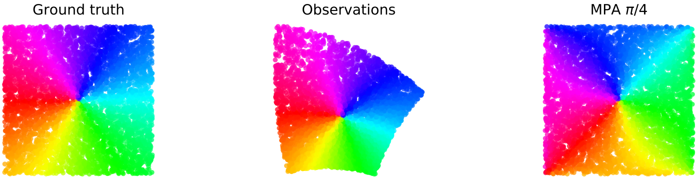
Another class of “spurious” solutions (Locatello et al., 2019):

$$\mathsf{a}^{\mathsf{R}_\theta}(p_{\mathsf{y}}) = \mathsf{F}_{\mathsf{y}}^{-1} \circ \boldsymbol{\Phi} \circ \mathsf{R}_\theta \circ \boldsymbol{\Phi}^{-1} \circ \mathsf{F}_{\mathsf{y}}$$

where F_{y} is the CDF of $\mathbf{y}$ and $\boldsymbol{\Phi}$ that of an isotropic Gaussian.

Given solution $\mathbf{y} = \mathbf{g}(\mathbf{x})$, yields *different* $\mathbf{y}'$ with same distribution by:

1. transforming $\mathbf{y}$ to an *isotropic* Gaussian via $\boldsymbol{\Phi}^{-1} \circ \mathsf{F}_{\mathsf{y}}$;
2. applying a rotation R_θ;
3. transforming back to the original distribution via $\mathsf{F}_{\mathsf{y}}^{-1} \circ \boldsymbol{\Phi}$.



Consequences for Unsupervised Representation Learning i

Spurious solutions vs. identifiability “except when”:

- Difference between the counterexamples we presented and the “exception” constituted by Gaussian components in the identifiability of linear ICA.
- In fact, both “spurious solutions” can be constructed for any choice of $p_{\mathbf{s}}$ with factorized components and invertible $\mathbf{f}$.
  - Unlike in linear ICA, where “spurious solutions” can only be constructed for Gaussian components (*corner case/exception*).

Implications for unsupervised machine learning:

- First implication: Nonlinear ICA is not $\sim_{\text{BSS}}$-identifiable, i.e., *nonlinear independent component estimation does not solve blind source separation.*

Consequences for Unsupervised Representation Learning ii

- **But there's a broader implication!**
- The Darmois construction can be applied even if the observed distribution is not generated by mixing independent components to begin with.
  - The converse implication is that any smooth distribution which is fully supported on $\mathbb{R}^n$ can be written as the mixture of n independent components through the Darmois construction.
- The Darmois construction can be used to show that unsupervised representation learning in the i.i.d. case (i.e., with any unconditional prior, see later) is impossible: **even if you assume a prior with non-independent components**.

Consequences for Unsupervised Representation Learning iii

- General prior on $\mathbf{z}$: it is enough to point out that we can transform any variable into independent (Darmois) Gaussian (CDF transform) variables; apply a rotation $\mathbf{O}$; then invert the Darmois construction;
- We get a nonlinear transformation $\mathbf{z}' = \mathbf{g}^{-1}(\mathbf{O}\mathbf{g}(\mathbf{z}))$ which has exactly the same distribution as $\mathbf{z}$ but is a complex nonlinear transformation thereof (individual components of $\mathbf{z}'$ could be mixtures of the components of $\mathbf{z}$).
- Take-home message: **By looking at the (i.i.d.) data alone and without further assumptions, it is not possible to recover the true latents, no matter what the prior may be**.
- Or in other terms: **Representations extracted by two different models fitting the data equally well may be arbitrarily entangled w.r.t. one another.**

Is All Hope Lost?

Identifiability in (nonlinear) representation learning: How?

- What about **assumptions on allowed source distributions** $\mathcal{P}$?
  - Failure of the naive approach for i.i.d. samples, regardless of $\mathcal{P}$;
  - Alternatives? → **Auxiliary Variable Setting**
- Another possible path could be to **constrain the class of the mixing functions** $\mathcal{F}$.
  - **Linear ICA:** If $\mathcal{F}$ is restricted to linear invertible matrices, we have $\sim_{\text{LIN}}$-identifiability.
  - Define a broader class of (nonlinear) functions s.t. interesting notion of identifiability can be achieved?

Nonlinear ICA with Auxiliary Variables

From Unconditional to Conditional Independence i

An *auxiliary variable* $\mathbf{u}$ renders the sources *conditionally* independent (Hyvärinen and Morioka, 2016, 2017; Hyvärinen et al., 2019).

$$\mathbf{s} \overset{\text{i.i.d.}}{\sim} p_{\mathbf{s}|\mathbf{u}}, \qquad p_{\mathbf{s}|\mathbf{u}}(\mathbf{s}|\mathbf{u}) = \prod_{i=1}^{n} p_{s_i|\mathbf{u}}(s_i|\mathbf{u}) \,. \tag{7}$$

With suitable assumptions, identifiability **without further restrictions on the nonlinear mixing f**, sometimes even up to $\sim_{\text{BSS}}$.

Why does it work?

- We showed that the Darmois construction provides a way to show nonidentifiability *even beyond the setting where $p_{\mathbf{s}} \in \mathcal{P}$ has independent components*.
- However, the Darmois construction relies on the assumption that the datapoints are (mixtures of) *independent and identically distributed* samples from a distribution $p_{\mathbf{s}}$, i.e., $\mathbf{s} \overset{\text{i.i.d.}}{\sim} p_{\mathbf{s}}$.

From Unconditional to Conditional Independence ii

- **It turns out that the auxiliary variable setting can in fact be interpreted as a deviation from the i.i.d. assumption.**
    - **Nonstationarity:** $\mathbf{u}$ indicates a time segment; *within* a time segment, the latent vector $\mathbf{s}$ is sampled i.i.d.; but the distribution changes *across* time segments, giving rise to nonstationarity.
    - **Autocorrelation:** stochastic process where $p_{\mathbf{s}(t)|\mathbf{s}(t-1)}$ has independent components. In this case, $\mathbf{u} = \mathbf{s}(t-1)$.
- In both cases, the variable $\mathbf{u}$ indicates **a deviation from the i.i.d. setting**: the $\mathbf{s}$ vectors may be

    (i) nonidentically distributed (in the nonstationary case their distribution changes depending on the time-segment label $\mathbf{u}$), or

    (ii) not independent (due to autocorrelation).

Generalised Contrastive Learning (Hyvärinen et al., 2019)

- A constructive proof of identifiability can be attained by exploiting contrastive learning.
- This technique transforms a density ratio estimation problem into one of supervised function approximation.
  - The auxiliary variables provide a handle to perform feature extraction (the aim is still representation learning).
- Generalized Contrastive Learning: Train a binary classifier. The two classes are

$$\text{Class 1} = (\mathrm{x}, \mathrm{u}); \qquad \text{Class 2} = (\mathrm{x}, \mathrm{u}^*)$$

- The feature extracted by the regression functions are the latent independent sources (up to tolerable ambiguities, under suitable assumptions on the regression function etc.)

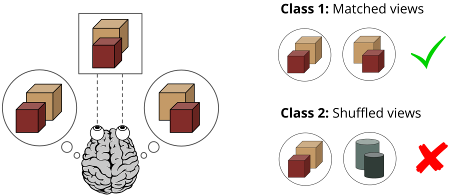
Example: Nonlinear ICA with Multiple Views (Gresele et al., 2019)

The Incomplete Rosetta Stone Problem: Two *sufficiently distinct* views, mixtures of element-wise independent corruptions of the same sources:

$$\mathbf{x}_1 = \mathbf{f}_1(\mathbf{s}); \qquad \mathbf{x}_2 = \mathbf{f}_2(\mathbf{s} + \mathbf{n})$$

$$p(\mathbf{s}) = \prod_i p(s_i); \qquad p(\mathbf{n}) = \prod_i p(n_i)$$

Learn to distinguish matched views from shuffled ones.



Class 1: Matched views ✓

Class 2: Shuffled views ✗

Under suitable assumptions, one can invert $\mathbf{f}_1$ and $\mathbf{f}_2$ (up to $\sim_{\text{BSS}}$).

Nonlinear ICA by Restricting the Function Class

Recovering Identifiability by Restricting the Function Class

- We talked about a space of models $\mathcal{M} \subseteq \mathcal{F} \times \mathcal{P}$;
- Instead of using auxiliary variables, one can stay in the i.i.d. setting and restrict the mixing function class $\mathcal{F}$.
- One example is linear ICA: the mixing function $\mathbf{f}$ is restricted to be linear.
- Other examples:
    - the post-nonlinear model (Taleb and Jutten, 1999; Zhang and Hyvärinen, 2009)
    - the minimal nonlinear distortion principle (Zhang and Chan, 2008)
    - conformal maps (Hyvärinen and Pajunen, 1999; Buchholz et al.)
- We (Gresele et al., 2021a) studied restrictions on the function class motivated by the **principle of independent causal mechanisms** (Schölkopf et al., 2012; Peters et al., 2017)

Independent mechanism analysis, a new concept?

Advances in Neural Information Processing Systems (NeurIPS), 2021

Luigi Gresele*,[1], Julius von Kügelgen*,[1,2], Vincent Stimper[1,2]
Bernhard Schölkopf[1], Michel Besserve[1]

* Equal contribution
[1] Empirical Inference Department, Max Planck Institute for Intelligent Systems, Tübingen
[2] Department of Engineering, University of Cambridge

Independent Causal Mechanisms and ICA

Research question

How can ideas from causality, particularly the principle of independent causal mechanisms, be useful for unsupervised representation learning tasks such as nonlinear ICA?

Principle: Independent Causal Mechanisms (ICM)

Principle of independent causal mechanisms (ICM)

"The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other". (Schölkopf et al., 2012; Peters et al., 2017).

Mechanisms/"autonomous modules" are *typically* understood as

- causal conditionals $P(V_i|\mathbf{PA}_i)$, or
- structural equations $V_i := f_i(\mathbf{PA}_i, U_i)$

Intuition: Nature chooses mechanisms independently of each other.

This is an informal statement: various formalisations exist, implementing different types of "non-statistical independence" (Janzing and Schölkopf, 2010; Janzing et al., 2010; Zscheischler et al., 2011; Shajarisales et al., 2015; Besserve et al., 2018; Janzing, 2021; Besserve et al., 2021)

Has proven useful for causal discovery & (non-iid) machine learning.

Principle: Independent Causal Mechanisms (ICM)

High-level intuition: No fine-tuning (Besserve et al., 2018).

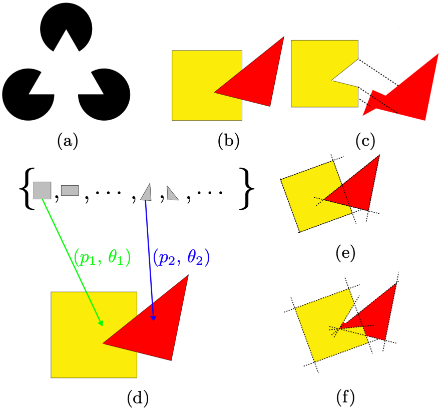
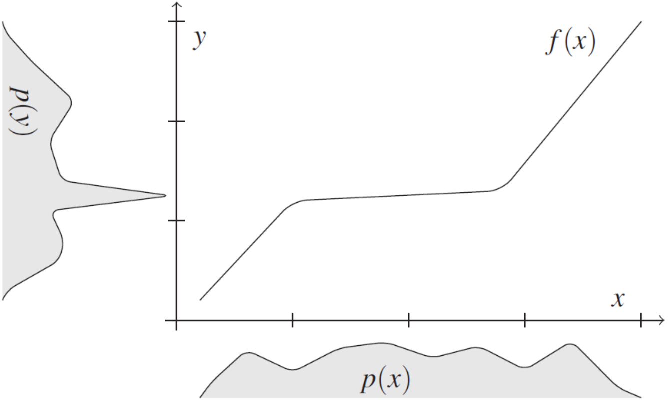


Figure (b) could be generated by a juxtaposition of shapes as in (c), but it would be a quite atypical realisation (requiring fine tuning).

A Visual Example of *Dependent* Input and Mechanism

For two variables, ICM implies independence between cause distribution $p(\mathbf{c})$ and mechanism $p(\mathbf{e}|\mathbf{c})$ (or $\mathbf{f}$).



Information-geometric causal inference (IGCI)

(figure from Janzing et al. (2012))

Classical Interpretation of ICM is *not* Useful for BSS

Naive approach: impose "independence" between the cause or source distribution $p_\mathbf{s}$ and the mechanism or mixing function $\mathbf{f}$.

Closest ICM criterion: Information-Geometric Causal Inference (IGCI) (Daniušis et al., 2010; Janzing et al., 2012) also assumes deterministic mapping between cause $\mathbf{s}$ and effect $\mathbf{x} = \mathbf{f}(\mathbf{s})$.

Postulates "independence" as lack of fine-tuning between $\mathbf{J_f}$ and $p_\mathbf{s}$:

$$\int \log |\mathbf{J_f}(\mathbf{s})| p(\mathbf{s}) d\mathbf{s} = \int \log |\mathbf{J_f}(\mathbf{s})| d\mathbf{s} \int p(\mathbf{s}) d\mathbf{s} = \int \log |\mathbf{J_f}(\mathbf{s})| d\mathbf{s}$$

Used for causal discovery where both cause and effect are observed.

Problem for use in BSS: Not invariant to reparametrisation, trivially satisfied when $p_\mathbf{s}$ is uniform (e.g., via the Darmois construction)

$\implies$ Independence between $p_\mathbf{s}$ and $\mathbf{f}$ is not sufficiently constraining when $p_\mathbf{s}$ is not observed (as in unsupervised learning).

Classical Interpretation of ICM is *not* Useful for BSS

Naive approach: impose “independence” between the cause or source distribution $p_{\mathbf{s}}$ and the mechanism or mixing function $\mathbf{f}$.

Closest ICM criterion: Information-Geometric Causal Inference (IGCI) (Daniušis et al., 2010; Janzing et al., 2012) also assumes deterministic mapping between cause $\mathbf{s}$ and effect $\mathbf{x} = \mathbf{f}(\mathbf{s})$.

Postulates “independence” as lack of fine-tuning between $\mathbf{J}_{\mathbf{f}}$ and $p_{\mathbf{s}}$:

$$\int \log |\mathbf{J}_{\mathbf{f}}(\mathbf{s})| p(\mathbf{s}) d\mathbf{s} = \int \log |\mathbf{J}_{\mathbf{f}}(\mathbf{s})| d\mathbf{s} \int p(\mathbf{s}) d\mathbf{s} = \int \log |\mathbf{J}_{\mathbf{f}}(\mathbf{s})| d\mathbf{s}$$

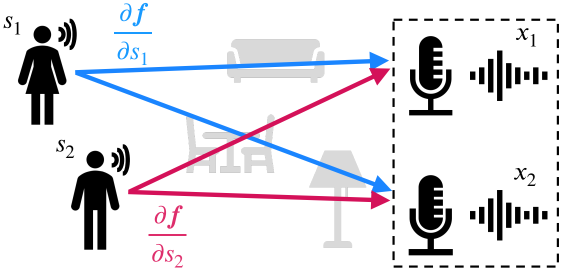
Used for causal discovery where both cause and effect are observed.

Problem for use in BSS: Not invariant to reparametrisation, trivially satisfied when $p_{\mathbf{s}}$ is uniform (e.g., via the Darmois construction)

$\implies$ Independence between $p_{\mathbf{s}}$ and $\mathbf{f}$ is not sufficiently constraining when $p_{\mathbf{s}}$ is not observed (as in unsupervised learning).

IMA and the Cocktail-Party Problem: Intuition

Postulate: the mechanisms by which each source influences the observed distribution should be “independent”.



s_1
$\frac{\partial f}{\partial s_1}$
s_2
$\frac{\partial f}{\partial s_2}$
x_1
x_2

Intuition: The location of the speakers and the room acoustics are not fine tuned to one another (violated, e.g., in a concert hall).

The *influences* of the individual sources on the observations are captured by the partial derivatives $\partial f / \partial s_i$.

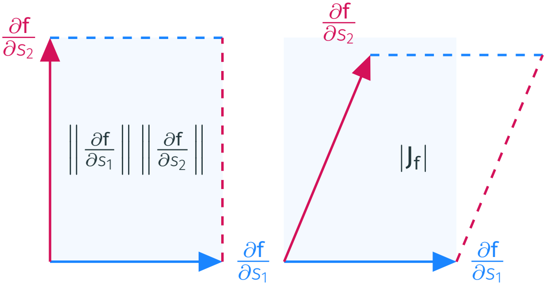
The Independent Mechanism Analysis Principle

Principle: independent mechanism analysis (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 $\mathbf{s}$:

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

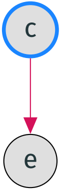
IMA is an orthogonality condition on the columns $\frac{\partial \mathbf{f}}{\partial s_i}$ of the Jacobian:


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

Comparison with IGCI

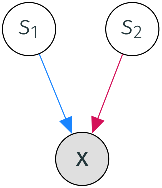
IGCI: decoupling of cause and mechanism,

$$\int \log|\mathrm{J_f}(\mathrm{s})|p(\mathrm{s})d\mathrm{s} = \int \log|\mathrm{J_f}(\mathrm{s})|d\mathrm{s} \int p(\mathrm{s})d\mathrm{s}$$



IMA: decoupling of the influence of each component,

$$\int \log|\mathrm{J_f}(\mathrm{s})|p(\mathrm{s})d\mathrm{s} = \int \sum_{i=1}^{n} \log\left|\left|\frac{\partial \mathrm{f}}{\partial s_i}\right|\right| p(\mathrm{s})d\mathrm{s}$$



Information-geometric interpretation of IMA: Causal effect of a soft intervention $\mathbf{s} \mapsto \boldsymbol{\sigma}(\mathbf{s})$ (KL between obs. and interv. dist.) decomposes as sum of effects of interventions on the individual sources.

The IMA Contrast C_{IMA} and its Properties

The IMA contrast C_{IMA}

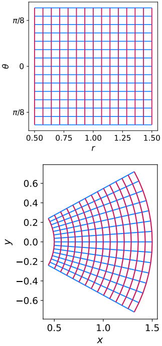
The IMA contrast $C_{\mathrm{IMA}}(\mathbf{f}, p(\mathbf{s}))$ is given by

$$C_{\mathrm{IMA}}(\mathbf{f}, p(\mathbf{s})) = \int \left(\sum_{i=1}^{n} \log \left\| \frac{\partial \mathbf{f}}{\partial s_i}(\mathbf{s}) \right\| - \log |\mathbf{J}_{\mathbf{f}}(\mathbf{s})| \right) p(\mathbf{s}) d\mathbf{s} \, .$$

Properties:

(i) $C_{\mathrm{IMA}} \geq 0$ with equality iff. $\mathbf{f}$ is an *orthogonal coordinate transformation*: $\mathbf{J}_{\mathbf{f}}(\mathbf{s}) = \mathbf{O}(\mathbf{s})\mathbf{D}(\mathbf{s})$.

(ii) C_{IMA} **is invariant to reparametrisation of the sources by permutation and element-wise transformation** (the irresolvable ambiguities of nonlinear BSS).



Q: Is IMA useful for identifiability? Can we use it to distinguish “spurious” solutions from the true one?

Theory: Darmois Solutions have Positive C_{IMA}

Theorem: C_{IMA} of the Darmois construction for nonlinear mixing

Assume that the data is generated by $\mathbf{x} = \mathbf{f}(\mathbf{s})\,,\; p(\mathbf{s}) = \prod_{i=1}^{n} p_i(s_i)$. Suppose that $x_i \not\perp\!\!\!\perp x_j$ for some $i \neq j$. Then $C_{\text{IMA}}^{\text{Darmois}} > 0$.

Consequence: for any $\mathbf{f}$ s.t. $C_{\text{IMA}}(\mathbf{f}, p(\mathbf{s})) = 0$, **we can distinguish the Darmois solution from the true one.**

Corollary: C_{IMA}-identifiability of conformal maps

If $\mathbf{f}$ is a conformal (angle-preserving) map, the true solution is distinguishable from Darmois-solutions via C_{IMA}.

Matches known result for $n = 2$ (Hyvärinen and Pajunen, 1999) and more recent result for general $n \geq 2$ (Buchholz et al., 2022).

Note: $C_{\text{IMA}} = 0$ for larger class of functions than just conformal maps.

Theory: Darmois Solutions have Positive C_{IMA}

Theorem: C_{IMA} of the Darmois construction for nonlinear mixing

Assume that the data is generated by $\mathbf{x} = \mathbf{f}(\mathbf{s})\,,\ p(\mathbf{s}) = \prod_{i=1}^{n} p_i(s_i)$. Suppose that $x_i \not\perp\!\!\!\perp x_j$ for some $i \neq j$. Then $C_{\text{IMA}}^{\text{Darmois}} > 0$.

Consequence: for any $\mathbf{f}$ s.t. $C_{\text{IMA}}(\mathbf{f}, p(\mathbf{s})) = 0$, **we can distinguish the Darmois solution from the true one.**

Corollary: C_{IMA}-identifiability of conformal maps

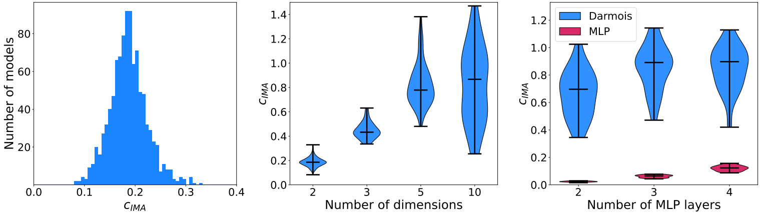
If $\mathbf{f}$ is a conformal (angle-preserving) map, the true solution is distinguishable from Darmois-solutions via C_{IMA}.

Matches known result for $n = 2$ (Hyvärinen and Pajunen, 1999) and more recent result for general $n \geq 2$ (Buchholz et al., 2022).

Note: $C_{\text{IMA}} = 0$ for larger class of functions than just conformal maps.

Numerical Simulations: C_{IMA} of the Darmois Construction

Data: $\mathbf{f}$ random Möbius transformation/MLP, $\mathbf{y} \sim U[0,1]^n$



(left) C_{IMA} for the Darmois construction is always larger than zero.

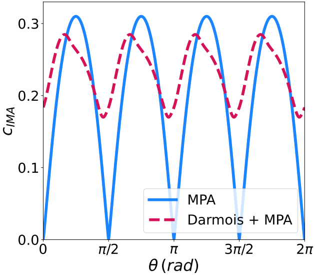
(center) Mean and variance grow with the dimensionality n.

(right) **Under deviations from our assumption** (L-layer MLP mixing), C_{IMA} **grows with** L**, but for the Darmois solution it is typically higher.**

Ruling out Measure-Preserving Automorphisms

Theorem: C_{IMA} of MPAs for conformal f

Assume that the data is generated by $\mathbf{x} = \mathbf{f}(\mathbf{s})\,,\;\; p(\mathbf{s}) = \prod_{i=1}^{n} p_i(s_i)$. Suppose that $\mathbf{f}$ is conformal and $\mathbf{s}$ non-Gaussian. Then $C_{\mathrm{IMA}} > 0$, unless $\mathbf{R}_\theta$ is "trivial" (i.e., a permutation matrix).



(blue) True solution composed with an MPA : C_{IMA} is periodic in θ; vanishes at multiples of $\pi/2$, i.e., when $\mathbf{R}_\theta$ is a permutation matrix.

(red) MPA composed with Darmois solution: $C_{\mathrm{IMA}} > 0 \;\forall\theta$.

Using C_{IMA} as a Learning Signal

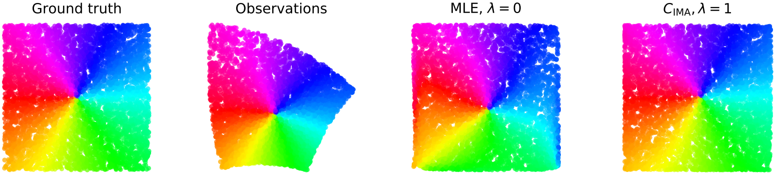
Regularised maximum-likelihood objective:

$$\mathcal{L}(\mathbf{g}; \mathbf{x}) = \mathbb{E}_{\mathbf{x}}[\log p_{\mathbf{g}}(\mathbf{x})] - \lambda\, C_{\mathrm{IMA}}(\mathbf{g}^{-1}, p_{\mathbf{y}})$$

where $\mathbf{g}$ is the learnt unmixing, $\mathbf{y} = \mathbf{g}(\mathbf{x})$ the reconstructed sources.

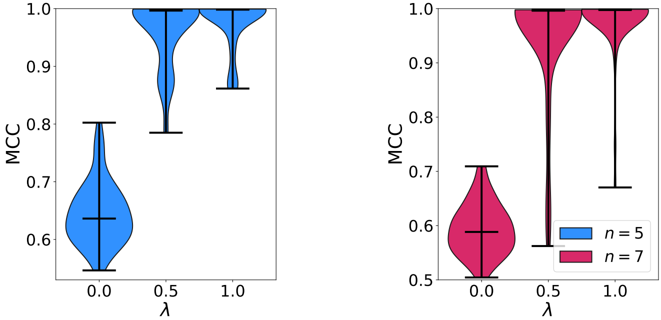
$\lambda = 0$: standard maximum likelihood estimation (MLE);
$\lambda > 0$: lower bound (exact iff. $C_{\mathrm{IMA}}(\mathbf{g}^{-1}, p_{\mathbf{y}}) = 0$).

Train residual flows (Chen et al., 2019) to maximise $\mathcal{L}$:



For $\lambda = 0$, learnt solutions do not achieve BSS; for $\lambda > 0$, they do.

IMA Helps To Solve Blind Source Separation



Mean correlation coefficient (MCC; higher is better) between true and reconstructed sources: $\lambda > 0$ is beneficial for BSS.

$\implies$ C_{IMA} **is a useful learning signal to recover the true solution.**

Summary

We proposed new ICM-inspired criterion for representation learning: independent mechanism analysis (IMA) postulates independence between the influences of individual sources; geometrically, orthogonality of $\frac{\partial \mathbf{f}}{\partial s_i}$.

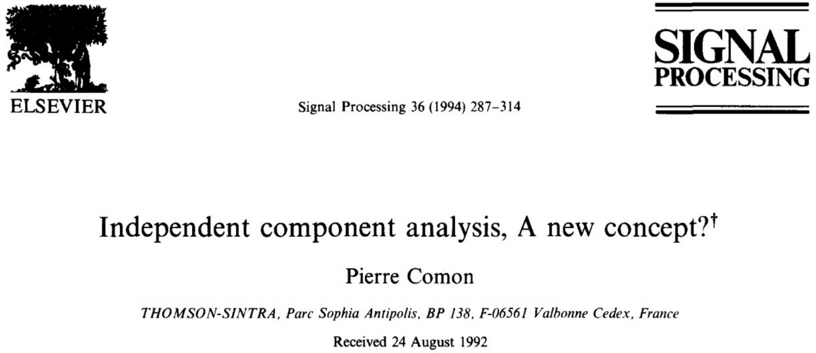
Proved that IMA rules out common "spurious" nonlinear ICA solutions; **more recently, Buchholz et al. (2022) proved *local identifiability of the IMA class*.**

- Full identifiability is still an open question.

Empirically, IMA-regularised maximum-likelihood learning helps to solve nonlinear blind source separation, **even under mild model misspecification**, as recently shown by Sliwa et al. (2022).

Paper: https://arxiv.org/abs/2106.05200
Code: https://github.com/lgresele/independent-mechanism-analysis

P.S. Explaining the Title:



ELSEVIER

Signal Processing 36 (1994) 287–314

SIGNAL PROCESSING

Independent component analysis, A new concept?†

Pierre Comon

THOMSON-SINTRA, Parc Sophia Antipolis, BP 138, F-06561 Valbonne Cedex, France

Received 24 August 1992

Embrace the Gap: VAEs Perform Independent Mechanism Analysis

Advances in Neural Information Processing Systems (NeurIPS), 2022

Patrik Reizinger*,[1], Luigi Gresele*,[2], Jack Brady*,[1], Julius von Kügelgen[2,3], Dominik Zietlow[2,4] Bernhard Schölkopf[2], Georg Martius[2], Wieland Brendel[1], Michel Besserve[2]

* Equal contribution
[1] University of Tübingen
[2] Max Planck Institute for Intelligent Systems, Tübingen
[3] Department of Engineering, University of Cambridge
[4] Amazon Web Services, Tübingen

Overview

- Variational autoencoders (VAEs) can be efficiently trained via variational inference by maximizing the evidence lower bound (ELBO), at the expense of a gap to the exact (log-)marginal likelihood.
- While some works claim that the ELBO is equal to the exact log-likelihood for near-deterministic decoders (Nielsen et al., 2020), other works claim that it includes inductive biases which are useful for representation learning (Rolinek et al., 2019; Kumar and Poole, 2020) (whereas maximizing the exact likelihood estimates a non-identifiable model).
- **Can we characterize the gap between ELBO and exact likelihood?**
- **Why would ELBO maximization yield useful representations?**

Deep Latent Variable Models (LVMs) (based on Rubenstein (2019))

- A Latent Variable Model (LVM) is a way to specify complex distribution over high dimensional spaces by composing simpler distributions (Bishop, 2006; Murphy, 2012).
- A LVM is specified by fixing
  - (i) a prior $p_{\mathrm{s}}(\mathrm{s})$;
  - (ii) a parametrized family of conditional distributions $p_{\boldsymbol{\theta}}(\mathrm{x}|\mathrm{s})$; also called the *decoder* or *generator* in the literature.
- Training a LVM requires
  - (a) picking a divergence between $p_{\boldsymbol{\theta}}(\mathrm{x})$ and the true data distribution $p_{\mathrm{data}}(\mathrm{x})$;
  - (b) finding (decoder) parameters $\boldsymbol{\theta}$ which minimize it.

Maximum Likelihood Training of Deep LVMs

- If we pick the Kullback-Leibler divergence:

$$\mathrm{KL}\left[p_{\text{data}}(\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{x})\right] = \mathbb{E}_{p_{\text{data}}}\left[\log p_{\text{data}}(\mathbf{x})\right] - \mathbb{E}_{p_{\text{data}}}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x})\right] . \quad (8)$$

- Since the left expectation doesn't depend on $\boldsymbol{\theta}$, minimizing the divergence is equivalent to maximizing the right expectation, which happens to be the log-likelihood.
- Optimizing the data likelihood $p_{\boldsymbol{\theta}}(\mathbf{x})$ in deep LVMs: Finding decoder parameters $\boldsymbol{\theta}$ maximizing

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})p_{\mathbf{s}}(\mathbf{s})d\mathbf{s}.$$

- Although $p_{\mathbf{s}}(\mathbf{s})$ and $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})$ are usually chosen to be simple and easy to evaluate, $p_{\boldsymbol{\theta}}(\mathbf{x})$ involves computing an integral.
- **This is difficult to evaluate in general, so approximate objectives are required.**

Variational Autoencoders (VAEs)

- Variational approximations (Struwe, 2000) introduce a family of conditional distributions called the *variational posterior*, $q_{\phi}(\mathsf{s}|\mathsf{x})$.
    - $q_{\phi}(\mathsf{s}|\mathsf{x})$ is a stochastic mapping $\mathsf{x} \mapsto \mathsf{s}$ with parameters $\boldsymbol{\phi}$.
- This allows to evaluate a **tractable lower bound** (Kingma and Welling, 2014; Rezende et al., 2014) of the model's log-likelihood termed **evidence lower bound (ELBO)**

$$\mathrm{ELBO}(\mathsf{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{q_{\phi}(\mathsf{s}|\mathsf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathsf{x}|\mathsf{s})\right] - \mathrm{KL}\left[q_{\phi}(\mathsf{s}|\mathsf{x}) || p_{\mathsf{s}}(\mathsf{s})\right]. \tag{9}$$

- The two terms in (9) are sometimes interpreted as a reconstruction term measuring the sample quality of the decoder and a regularizer—the Kullback-Leibler Divergence (KL) between the prior and the encoder (Kingma and Welling, 2019).

The ELBO Gap

- The variational approximation trades off computational efficiency with a difference w.r.t. the exact log-likelihood;
- This can be expressed alternatively as (see Doersch (2016); Kingma and Welling (2019))

$$\mathrm{ELBO}(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \underbrace{\log p_{\boldsymbol{\theta}}(\mathbf{x})}_{\text{(i)}} - \underbrace{\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x}) || p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})\right]}_{\text{(ii)}}, \tag{10}$$

where:

(i) is the *exact log-likelihood under our model $\boldsymbol{\theta}$;*

(i) is the KL-divergence between variational and true posteriors: a.k.a. the *ELBO gap*.

If the variational family of $q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})$ does not include $p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})$, the evidence lower bound (ELBO) will be strictly smaller than the exact likelihood $\log p_{\boldsymbol{\theta}}(\mathbf{x})$.

Common Modelling Choices in VAEs

- Variational Autoencoders (VAEs) rely on the variational approximation we described to train deep Latent Variable Models (LVMs);
- Deep neural networks parametrize the *encoder* $q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})$ and the *decoder* $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})$.
- Some common modelling choices (which we assume throughout in our work):
  - **The variational family of $q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})$ is a factorized Gaussian** with posterior means $\mu_k^{\boldsymbol{\phi}}(\mathbf{x})$, and variances $\sigma_k^{\boldsymbol{\phi}}(\mathbf{x})^2$ for the k^{th} factor $s_k|\mathbf{x}$, and with a diagonal covariance

$$s_k|\mathbf{x} \sim \mathcal{N}(\mu_k^{\boldsymbol{\phi}}(\mathbf{x}), \sigma_k^{\boldsymbol{\phi}}(\mathbf{x})^2)\,; \qquad (11)$$

  - **The decoder is a factorized Gaussian**, conditional on $\mathbf{s}$, with mean $\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s})$ and an isotropic covariance in n dimensions,

$$\mathbf{x}|\mathbf{s} \sim \mathcal{N}\left(\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}), \gamma^{-2}\mathbf{I}_n\right). \qquad (12)$$

The Deterministic Limit of VAEs i

- The stochasticity of VAEs makes it nontrivial to relate them to generative models with deterministic decoders such as those used in normalizing flows or independent component analysis.
- Some previous works postulate a *deterministic regime* (where the decoder precision γ^2 becomes infinite).
- Nielsen et al. (2020) explored this deterministic limit and argued that *deterministic* VAEs optimize an exact log-likelihood, similar to normalizing flows (Rezende and Mohamed, 2015; Papamakarios et al., 2019).
- The likelihood of the original variables becomes

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \log p_{\mathbf{s}}(\mathbf{s}) - \log \left| \mathrm{J}_{\mathbf{f}^{\boldsymbol{\theta}}} (\mathbf{s}) \right| . \qquad (13)$$

The Deterministic Limit of VAEs ii

- The comparison is nontrivial: VAEs contain an encoder and a decoder, whereas the likelihood for normalizing flows is written in terms of a single architecture.
- Nielsen et al. (2020) made this analogy by resorting to what we call a *self-consistency assumption*, stating that the VAE encoder inverts the decoder: **This is assumed but not proved.**

Definition (Near-deterministic self-consistency)

For a fixed $\boldsymbol{\theta}$, assume that mean decoder $\mathbf{f}^{\boldsymbol{\theta}}$ is invertible with inverse $\mathbf{g}^{\boldsymbol{\theta}}$, and that a map associates each choice of decoder parameters and observation $(\boldsymbol{\theta}, \gamma, \mathbf{x})$ to an encoder parameter $(\boldsymbol{\theta}, \gamma, \mathbf{x}) \mapsto \widehat{\phi}(\boldsymbol{\theta}, \gamma, \mathbf{x})$, we say the VAE is self-consistent whenever

$$\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x}) \to \mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) \quad \text{and} \quad \boldsymbol{\sigma}^{\widehat{\phi}}(\mathbf{x})^2 \to \mathbf{0}\ , \text{as } \gamma \to +\infty\,. \tag{14}$$

Assumptions

Assumption (Factorized VAE class with isotropic Gaussian decoder and log-concave prior)

We are given a fixed latent prior and three parameterized classes of $\mathbb{R}^d \to \mathbb{R}^d$ *mappings: the mean decoder class* $\boldsymbol{\theta} \mapsto \mathbf{f}^{\boldsymbol{\theta}}$*, and the mean and standard deviation encoder classes,* $\boldsymbol{\phi} \mapsto \boldsymbol{\mu}^{\boldsymbol{\phi}}$ *and* $\boldsymbol{\phi} \mapsto \boldsymbol{\sigma}^{\boldsymbol{\phi}}$ *s.t.*

(i) $p_{\mathbf{s}}(\mathbf{s}) \sim \prod_k m(s_k)$*, with* m *being smooth and fully supported on* $\mathbb{R}$*, having bounded non-positive second-order, and bounded third-order logarithmic derivatives;*

(ii) *the encoder and decoder are of the form in* (11) *and* (12)*, with isotropic decoder covariance* $1/\gamma^2 \mathbf{I}_n$*;*

(iii) *the variational mean and variance encoder classes are universal approximators;*

(iv) *for all* $\boldsymbol{\theta}$*,* $f^{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^d$ *is a bijection with inverse* $\mathbf{g}^{\boldsymbol{\theta}}$*, and both are* C^2 *with bounded first and second order derivatives.*

The Self-Consistent ELBO

- An obstacle in the theoretical analysis of VAEs is the unknown behaviour of the encoder unknown behaviour of the encoder

$$\mathrm{ELBO}(\mathsf{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \log p_{\boldsymbol{\theta}}(\mathsf{x}) - \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathsf{s}|\mathsf{x}) || p_{\boldsymbol{\theta}}(\mathsf{s}|\mathsf{x})\right], \tag{15}$$

- We will assume “the encoder does what it should” → minimize the gap at fixed decoder.
- Any associated optimal choice of encoder parameters satisfies

$$\widehat{\boldsymbol{\phi}}(\mathsf{x}, \boldsymbol{\theta}) \in \arg\max_{\boldsymbol{\phi}} \mathrm{ELBO}(\mathsf{x}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \arg\min_{\boldsymbol{\phi}} \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathsf{s}|\mathsf{x}) || p_{\boldsymbol{\theta}}(\mathsf{s}|\mathsf{x})\right].$$

- We call self-consistent ELBO the resulting

$$\mathrm{ELBO}^*(\mathsf{x}; \boldsymbol{\theta}) = \mathrm{ELBO}(\mathsf{x}; \boldsymbol{\theta}, \widehat{\boldsymbol{\phi}}(\mathsf{x}, \boldsymbol{\theta})). \tag{16}$$

- While in VAEs encoder and decoder are optimized jointly, we argue that for infinite capacity decoders

$$\underset{\boldsymbol{\theta}, \boldsymbol{\phi}}{\text{maximize}}\, \mathbb{E}_{\mathsf{x}\sim p(\mathsf{x})}\left[\mathrm{ELBO}(\mathsf{x}; \boldsymbol{\theta}, \boldsymbol{\phi})\right] \iff \underset{\boldsymbol{\theta}}{\text{maximize}}\, \mathbb{E}_{\mathsf{x}\sim p(\mathsf{x})}\left[\mathrm{ELBO}^*(\mathsf{x}; \boldsymbol{\theta})\right]$$

Proving Self-Consistency

- Self-consistent behaviour of the optimal choice of encoder parameters $\widehat{\phi}(\mathrm{x}, \boldsymbol{\theta}) \in \arg\max_{\boldsymbol{\phi}} \mathrm{ELBO}(\mathrm{x}; \boldsymbol{\theta}, \boldsymbol{\phi})$:

Proposition (Self-consistency of near-deterministic VAEs)

Under our assumptions, for all x, $\boldsymbol{\theta}$, *as* $\gamma \to +\infty$, *there exists at least one global minimum solution for* $\widehat{\boldsymbol{\phi}}(\mathrm{x}, \boldsymbol{\theta})$. *These solutions satisfy*

$$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathrm{x}) = \mathrm{g}^{\boldsymbol{\theta}}(\mathrm{x}) + O(1/\gamma) \quad \textit{and} \quad \sigma_k^{\widehat{\boldsymbol{\phi}}}(\mathrm{x})^2 = O(1/\gamma^2)\,,\ \textit{for all } k\,. \tag{17}$$

- In the limit of large γ, minimizing the ELBO gap (equivalently, maximizing the ELBO) w.r.t. the encoder parameters implies that the (mean) encoder “inverts” the (mean) decoder.

Near-Deterministic VAEs Perform IMA

Theorem (VAEs with a near-deterministic decoder approximate the Independent Mechanism Analysis (IMA) objective)

Under our assumptions, the variational posterior satisfies

$$\sigma_k^{\widehat{\phi}}(\mathrm{x})^2 = \left(-\frac{d^2 \log p_0}{ds_k^2}(g_k^{\boldsymbol{\theta}}(\mathrm{x})) + \gamma^2 \left\| \left[\mathrm{J}_{\mathbf{f}^{\boldsymbol{\theta}}}\left(\mathbf{g}^{\boldsymbol{\theta}}(\mathrm{x})\right)\right]_{:k} \right\|^2\right)^{-1} + O(1/\gamma^3)\,, \tag{18}$$

and the self-consistent ELBO *approximates the* IMA*-regularized log-likelihood:*

$$\mathrm{ELBO}^*(\mathrm{x};\boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\mathrm{x}) - c_{\mathrm{IMA}}(\mathbf{f}^{\boldsymbol{\theta}}, \mathbf{g}^{\boldsymbol{\theta}}(\mathrm{x})) + O_{\gamma\to\infty}\left(1/\gamma^2\right). \tag{19}$$

- Where c_{IMA} is given by

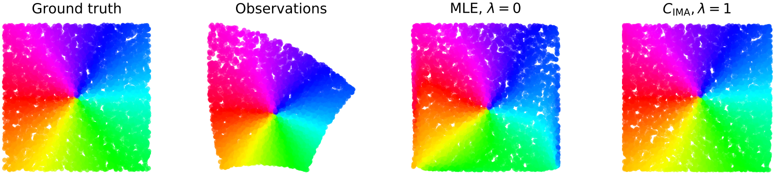
$$c_{\mathrm{IMA}}(\mathbf{f}^{\boldsymbol{\theta}}, \mathrm{s}) = \sum_{k=1}^{d} \log \left\| \frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial s_k}(\mathrm{s}) \right\| - \log \left| \mathrm{J}_{\mathbf{f}^{\boldsymbol{\theta}}}(\mathrm{s}) \right|$$

IMA, C_{IMA} & Identifiability

- c_{IMA} measures how much the Jacobian columns deviate from being orthogonal; $C_{\text{IMA}} = 0$ defines the IMA function class.
- IMA was motivated by identifiability (Gresele et al., 2021b); Recent advances (Buchholz et al., 2022) show *local identifiability*.
- In Gresele et al. (2021b), we used a regularized likelihood objective to estimate the IMA model:

$$\mathcal{L}_{\text{IMA}}(\mathbf{g}; \mathbf{x}) = \mathbb{E}_{\mathbf{x}}[\log p_{\mathbf{g}}(\mathbf{x})] - \lambda\, C_{\text{IMA}}(\mathbf{g}^{-1}, p_{\mathbf{y}})$$

  where $\mathbf{g}$ is the learnt decoder, $\mathbf{y} = \mathbf{g}(\mathbf{x})$ the reconstructed sources. $\lambda = 0$: standard maximum likelihood estimation (MLE); $\lambda > 0$: lower bound (exact iff. $C_{\text{IMA}}(\mathbf{g}^{-1}, p_{\mathbf{y}}) = 0$).



Asymptotic Self-Consistent Behaviour

- Self-consistency holds in practical VAE training.

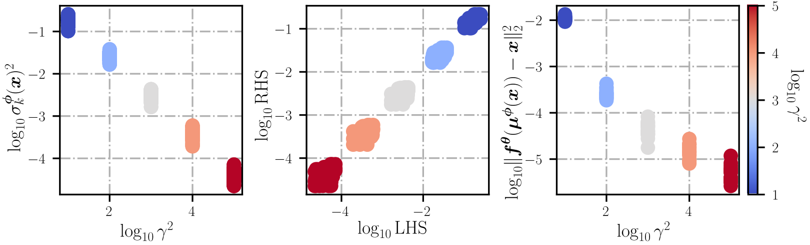


Figure 10: Self-consistency in *VAE* training, on a log-log plot. **Left**: convergence of $\sigma_k^{\widehat{\phi}}(\mathrm{x})^2$ to 0; **Center:** connecting $\sigma_k^{\widehat{\phi}}(\mathrm{x})^2$, γ^2, and the column norms of the decoder Jacobian via LHS and RHS of (18); **Right:** convergence of $\boldsymbol{\mu}^{\widehat{\phi}}(\mathrm{x})$ to $\mathbf{g}^{\boldsymbol{\theta}}$ (x)

Visualising the ELBO Gap

- We generate some data by the nonlinear ICA generative model, and fix the VAE decoder to true mixing (optimal decoder; with change of variables, it gives the true data likelihood).
- We train the encoder and plot the ELBO*, the IMA-regularized and unregularized log-likelihoods for different γ^2.

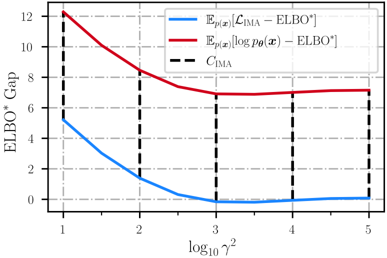


Figure 11: With a decoder outside the IMA-class, i.e., a decoder with $C_{\text{IMA}} > 0$.

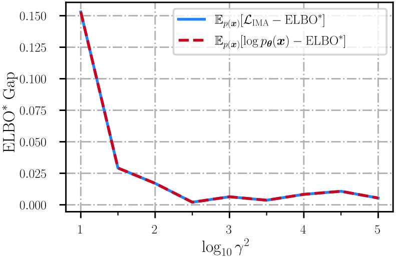


Figure 12: With an IMA-class decoder, i.e., a decoder with $C_{\text{IMA}} = 0$.

Identifiability

- When we generate data according to the IMA model, VAEs successfully separate the sources (i.e., they solve blind source separation).

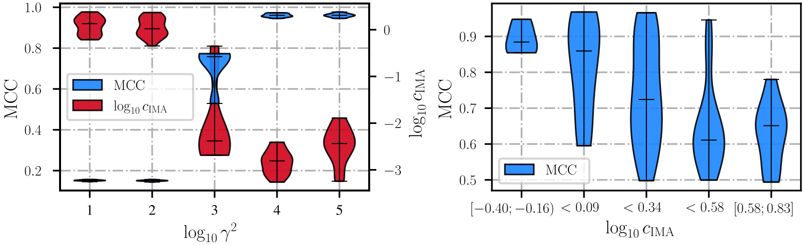


Figure 13: Left: c_{IMA} and Mean Correlation Coefficient (MCC) for 3-dimensional Möbius mixings **Right:** MCC depending on the *volume-preserving linear map's* c_{IMA} ($\gamma^2 = 1\mathrm{e}5$)

Disentanglement on a Simple Image Dataset

- We test VAEs ability to disentangle on a simple image dataset;
- Our choice is motivated by (Horan et al., 2021; Donoho and Grimes, 2005) showing that this data-generating process may approximately satisfy the IMA principle.

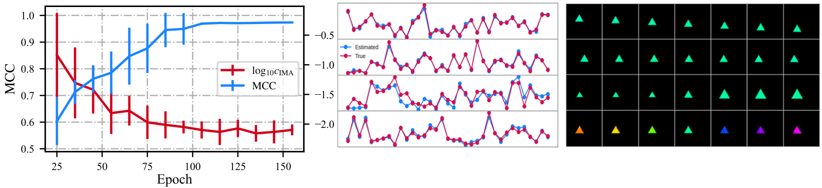


Figure 14: **Left:** c_{IMA} and MCC for Sprites (Watters et al., 2019) during training ($\gamma^2 = 1$); **Center:** true and estimated latent factors for the best trained VAE on Sprites; **Right:** the corresponding latent interpolations and MCC values (from top to bottom): *y*- (0.989), *x*-position (0.996), scale (0.933), and color (0.989).

Discussion

- Identifiability (=disentanglement) of variational generative models can be studied by leveraging result from the field of ICA,
- Design of encoder an decoder model class → inductive bias of the gap.
- We characterize the ELBO gap in the near-deterministic regime.
- Limitations:
  - near-determinism is computationally challenging;
  - non-amortized inference;
  - same latent and observation dimensions;
  - generalization to β-VAEs: similar under self-consistency assumptions.

Paper: https://arxiv.org/abs/2206.02416
Code: https://github.com/rpatrik96/ima-vae

Conclusion

Identifiability in Representation Learning & Causal Inference i

In a recent interview, Judea Pearl described his contribution to causality as follows (Pearl, 2022):

> ***I have focused on the problem of identification, rather than estimation.*** *This calls for transforming the desired causal quantity into an equivalent probabilistic expression (called estimand) that can be estimated from data. Once an estimand is derived, the actual estimation step is no longer causal, and can be accomplished by standard statistical methods. This is indeed where machine learning excels, unlike the identification step in which machine learning and standard statistical methods are almost helpless. It is for this reason that I focus on identification – this is where the novelty of causal thinking lies, and where a new calculus had to be developed.*

Identifiability in Representation Learning & Causal Inference ii

The focus on identifiability is a common theme between causal inference and nonlinear ICA (Hyvärinen et al., 2019):

> *The essential difference [between nonlinear ICA and] most methods for unsupervised representation learning is that the approach starts by defining a generative model in which the original latent variables can be recovered, i.e. the model is identifiable by design.*

ICA for Causal Inference (& Causality for representation learning?):

- ICA can be used as a method for *causal discovery*.
- Methods based on ICA allow going beyond the Markov equivalence class of the true DAG—subject to additional assumptions, mirroring those used for identifiability of ICA (Shimizu et al., 2006; Monti et al., 2020; Khemakhem et al., 2021).

Thank you for your attention!

Happy to take any questions

References

References

P. Ablin. *Exploration of multivariate EEG /MEG signals using non-stationary models*. PhD thesis, 2019. Thèse de doctorat dirigée par Cardoso, Jean-François et Gramfort, Alexandre Mathématiques et Informatique Université Paris-Saclay (ComUE) 2019.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

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.

M. Besserve, R. Sun, D. Janzing, and B. Schölkopf. A theory of independent mechanisms for extrapolation in generative models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6741–6749, 2021.

C. M. Bishop. *Pattern Recognition and Machine Learning*, volume 4. Springer New York, 2006. doi: 10.1007/978-0-387-45528-0. URL https://doi.org/10.1007/978-0-387-45528-0.

S. Buchholz, M. Besserve, and B. Schölkopf. Function classes for identifiable nonlinear independent component analysis. In *UAI 2022 Workshop on Causal Representation Learning*.

S. Buchholz, M. Besserve, and B. Schölkopf. Function Classes for Identifiable Nonlinear Independent Component Analysis. *Advances in Neural Information Processing Systems*, 36, 2022.

R. T. Chen, J. Behrmann, D. Duvenaud, and J.-H. Jacobsen. Residual flows for invertible generative modeling. *arXiv preprint arXiv:1906.02735*, 2019.

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

P. Daniušis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 143–150, 2010.

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

C. Doersch. Tutorial on Variational Autoencoders. *ArXiv preprint*, abs/1606.05908, 2016.

D. L. Donoho and C. Grimes. Image Manifolds which are Isometric to Euclidean Space. *J. Math. Imaging Vis.*, 23(1):5–24, July 2005. ISSN 0924-9907, 1573-7683. doi: 10.1007/s10851-005-4965-4. URL https://doi.org/10.1007/s10851-005-4965-4.

L. Gresele, P. K. Rubenstein, A. Mehrjou, F. Locatello, and B. Schölkopf. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. In *Uncertainty in Artificial Intelligence*, pages 217–227. PMLR, 2019.

L. Gresele, J. Von Kügelgen, V. Stimper, B. Schölkopf, and M. Besserve. Independent mechanism analysis, a new concept? *Advances in Neural Information Processing Systems*, 34, 2021a.

L. Gresele, J. von Kügelgen, V. Stimper, B. Schölkopf, and M. Besserve. Independent mechanisms analysis, a new concept? In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 28233–28248. Curran Associates, Inc., Dec. 2021b. URL https://proceedings.neurips.cc/paper/2021/file/edc27f139c3b4e4bb29d1cdbc45663f9-Paper.pdf.

G. E. Hinton. Learning multiple layers of representation. *Trends in cognitive sciences*, 11(10): 428–434, 2007.

D. Horan, E. Richardson, and Y. Weiss. When is unsupervised disentanglement possible? *Adv. Neur. In.*, 34, 2021.

F. Huszár. Goals and principles of representation learning, 4 2018.

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

A. Hyvärinen and H. Morioka. Nonlinear ICA of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, Ltd, 2001.

A. Hyvärinen, H. Sasaki, and R. 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.

D. Janzing. Causal version of principle of insufficient reason and maxent. *arXiv preprint arXiv:2102.03906*, 2021.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

D. Janzing, P. O. Hoyer, and B. Schölkopf. Telling cause from effect based on high-dimensional observations. In *International Conference on Machine Learning*, 2010.

D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

I. Khemakhem, D. Kingma, R. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020a.

I. Khemakhem, R. Monti, D. Kingma, and A. Hyvarinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. *Advances in Neural Information Processing Systems*, 33:12768–12778, 2020b.

I. Khemakhem, R. Monti, R. Leech, and A. Hyvarinen. Causal autoregressive flows. In *International Conference on Artificial Intelligence and Statistics*, pages 3520–3528. PMLR, 2021.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In Y. Bengio and Y. LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

D. P. Kingma and M. Welling. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000056. URL https://doi.org/10.1561/2200000056. arXiv: 1906.02691.

A. Kumar and B. Poole. On Implicit Regularization in β-VAEs. In *International Conference on Machine Learning*, pages 5480–5490. PMLR, 2020.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

D. Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.

R. P. Monti, K. Zhang, and A. Hyvärinen. Causal discovery with general non-linear relationships using non-linear ICA. In *Uncertainty in Artificial Intelligence*, pages 186–195. PMLR, 2020.

L. Moschella, V. Maiorca, M. Fumero, A. Norelli, F. Locatello, and E. Rodolà. Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430*, 2022.

K. P. Murphy. *Machine learning: A probabilistic perspective*. MIT press, 2012.

D. Nielsen, P. Jaini, E. Hoogeboom, O. Winther, and M. Welling. SurVAE flows: Surjections to bridge the gap between VAEs and flows. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H.-T. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *ArXiv preprint*, abs/1912.02762, 2019.

J. Pearl. Interview with judea pearl. *Observational Studies*, 8(2):23–36, 2022.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. In F. R. Bach and D. M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1530–1538. JMLR.org, 2015.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1278–1286. JMLR.org, 2014.

G. Roeder, L. Metz, and D. Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pages 9030–9039. PMLR, 2021.

M. Rolinek, D. Zietlow, and G. Martius. Variational autoencoders pursue PCA directions (by accident). In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12406–12415. IEEE, June 2019. doi: 10.1109/cvpr.2019.01269. URL *https://doi.org/10.1109/cvpr.2019.01269*.

P. K. Rubenstein. Variational Autoencoders are not autoencoders, Jan 2019. URL *http://paulrubenstein.co.uk/variational-autoencoders-are-not-autoencoders/*.

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.

N. Shajarisales, D. Janzing, B. Schoelkopf, and M. Besserve. Telling cause from effect in deterministic linear dynamical systems. In *International Conference on Machine Learning*, pages 285–294. PMLR, 2015.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

V. Skitović. On a property of a normal distribution. In *Doklady Akad. Nauk*, 1953.

J. Sliwa, S. Ghosh, V. Stimper, L. Gresele, and B. Schölkopf. Probing the robustness of independent mechanism analysis for representation learning. In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

M. Struwe. *Variational Methods*, volume 991. Springer Berlin Heidelberg, 2000. ISBN 9783662041963, 9783662041949. doi: 10.1007/978-3-662-04194-9. URL *https://doi.org/10.1007/978-3-662-04194-9*.

A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *IEEE Transactions on signal Processing*, 47(10):2807–2820, 1999.

V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

L. Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.

N. Watters, L. Matthey, S. Borgeaud, R. Kabra, and A. Lerchner. Spriteworld: A flexible, configurable reinforcement learning environment. https://github.com/deepmind/spriteworld/, 2019. URL *https://github.com/deepmind/spriteworld/*.

J. Wu, Y. Belinkov, H. Sajjad, N. Durrani, F. Dalvi, and J. Glass. Similarity analysis of contextual word representation models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4638–4655, 2020.

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

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 647–655. AUAI Press, 2009.

J. Zscheischler, D. Janzing, and K. Zhang. Testing whether linear equations are causal: A free probability theory approach. In *27th Conference on Uncertainty in Artificial Intelligence*, pages 839–847, 2011.