The Incomplete Rosetta Stone Problem: Identifiability Results for Multi-View Nonlinear ICA

Luigi Gresele[*1,2], **Paul K. Rubenstein**[*1,3], **Arash Mehrjou**[1,4], **Francesco Locatello**[1,5] **and Bernhard Schölkopf**[1]
[1]Empirical Inference, Max Planck Institute for Intelligent Systems, Tübingen, Germany.
[2]Max Planck Institute for Biological Cybernetics, Tübingen, Germany.
[3]Machine Learning Group, University of Cambridge, United Kingdom.
[4]Max Planck ETH Center for Learning Systems, Zürich, Switzerland.
[5]BMI, Dept. for Computer Science, ETH Zürich, Switzerland.

Abstract

We consider the problem of recovering a common latent source with independent components from multiple views. This applies to settings in which a variable is measured with multiple experimental modalities, and where the goal is to synthesize the disparate measurements into a single unified representation. We consider the case that the observed views are a nonlinear mixing of component-wise corruptions of the sources. When the views are considered separately, this reduces to nonlinear Independent Component Analysis (ICA) for which it is provably impossible to undo the mixing. We present novel identifiability proofs that this is possible when the multiple views are considered jointly, showing that the mixing can theoretically be undone using function approximators such as deep neural networks. In contrast to known identifiability results for nonlinear ICA, we prove that independent latent sources with arbitrary mixing can be recovered as long as multiple, sufficiently different noisy views are available.

1 INTRODUCTION

We consider the setting described by the following generative model

$$\boldsymbol{x}_1 = \boldsymbol{f}_1(\boldsymbol{s}) \tag{1}$$

$$\boldsymbol{x}_2 = \boldsymbol{f}_2(\boldsymbol{s}) \tag{2}$$

$$p(\boldsymbol{s}) = \prod_i p_i(s_i)\,, \tag{3}$$

where $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{s} \in \mathbb{R}^D$ and $\boldsymbol{f}_1, \boldsymbol{f}_2$ are arbitrary smooth and invertible transformations of the latent variable $\boldsymbol{s} = (s_1, \dots, s_D)$ with mutually independent components. The goal is to recover $\boldsymbol{s}$, undoing the mixing induced by the $\boldsymbol{f}_i$, in the case where only observations of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are available.

The two decoupled problems defined by considering pairs of Equations 1, 3 and 2, 3 separately are instances of Independent Component Analysis (ICA). This unsupervised learning method aims at providing a principled approach to disentanglement of independent latent components, blind source separation, and feature extraction [21]. Its applications are ubiquitous, including neuroimaging [28], signal processing [34], text mining [17], astronomy [31] and financial time series analysis [32]. An ICA problem is identifiable when it is provably possible to simultaneously undo the mixing and recover the sources $\boldsymbol{s}$ up to tolerable ambiguities. Proofs of identifiability are crucial for the characterization of reliable ICA methods; in absence of these, we cannot be confident that a method successfully retrieves the true sources, even within controlled settings.

The case in which $\boldsymbol{f}_i$ is a linear function, called linear ICA, has been shown to be identifiable if at most one of the latent components is Gaussian [10, 38, 9]. This triggered the development of algorithms and encouraged their application. In contrast, the nonlinear ICA problem was shown to be provably unidentifiable without further assumptions on the data generating process [22]. Much research in this field has thus attempted to characterize the assumptions under which identifiability holds. Such assumptions may be grouped into two main categories: (i) those regarding properties of the sources (e.g. non-stationarity or time correlation in time series settings [6, 37]); and (ii) those restricting the functional form of the mixing functions (e.g., post-nonlinear mixing [40]).

A recent breakthrough was to leverage a technique known as contrastive learning, a method recasting

[*]Equal contribution.

the problem of unsupervised learning as a supervised one [15, 19, 20, 23]. This is a powerful proof technique, which additionally provides algorithms which can be practically implemented using modern deep learning frameworks. The setup in [19, 20, 23] makes strong assumptions on the data generating mechanism, but allows for arbitrary nonlinear mixing of the sources. However, the unconditional independence assumption of the sources (Equation 3) is replaced by a *conditional* independence statement, and requires observations of the additional variable conditioned on.

In this paper, we employ contrastive learning to address the setting specified by Equations 1–3, where in contrast to [23], no observations of parent variables of the sources are available. This corresponds to cases in which multiple recordings of the same process, acquired with different instruments and possibly different modalities, are available, and the goal is to find an unambiguous representation of the latent state common to all. Multiview settings of this sort are common in large biomedical and neuroimaging datasets [1, 30, 41, 36], motivating the need for reliable statistical tools enabling simultaneous handling of multiple sets of variables.

As a metaphor for such a setting, consider the story of the Rosetta Stone, a stele discovered during Napoleon's campaign in Egypt in 1799, inscribed with three versions of a decree issued at Memphis in 196 BC. The realization that the stone reported the same text translated into three different languages led the French philologist Champollion to succeed in translating two unknown languages (Ancient Egyptian, in hieroglyphic script and Demotic script) by exploiting a known one (Ancient Greek). Rather, we consider the radically unsupervised task in which, given a Rosetta Stone with only two texts, both in unknown languages, we want to learn an unambiguous common representation for both of them.

The main contribution of this paper is to show that jointly addressing multiple demixing problems allows for identifiability with assumptions which do not directly refer to the sources, nor to restriction of the class of mixing functions, but rather to the conditional probability distribution of one observation given the other. This provides identifiability results in a novel setting, with assumptions entailing a different interpretation - namely, that the views have to be sufficiently diverse.

The remainder of this paper is organized as follows. In Section 2 we provide background information about the technique of contrastive learning for ICA and briefly review recent work that employs it. In Section 3 we present our main results, providing identifiability for different multi-view settings. In Section 4 we discuss other relevant works in the literature. Finally, we summarize and discuss our results in Section 5.

2 NONLINEAR ICA WITH CONTRASTIVE LEARNING

Consider the nonlinear ICA setting, where observations of a variable $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{s})$ are available, where $\boldsymbol{f}$ is an arbitrary nonlinear invertible mixing. The proof of non-identifiability for the general case with unconditionally independent sources was an important negative result [22]. We review it briefly in Appendix A.

A proposed modification of this setting [23] involves an auxiliary observed variable $\boldsymbol{u}$ and a change in the independence properties. If the *unconditional independence* is substituted with a *conditional independence* given the auxiliary variable $\boldsymbol{u}$, i.e.

$$\log p(\boldsymbol{s}|\boldsymbol{u}) = \sum_i q_i(s_i, \boldsymbol{u}) \,, \tag{4}$$

for some functions q_i, the model becomes identifiable. The conditional independence statement in Equation 4 can be interpreted as positing that $\boldsymbol{u}$ is a parent of the sources $\boldsymbol{s}$. A further assumption on the effect of variations in $\boldsymbol{u}$ on $\boldsymbol{x}$, called *variability* in the paper, is required. Intuitively, it demands that $\boldsymbol{u}$ has a sufficiently diverse influence on $\boldsymbol{x}$.

In the setting described above, a constructive proof of identifiability is attained by exploiting contrastive learning [15]. This technique transforms a density ratio estimation problem into one of supervised function approximation. This idea has a long history [12], and has attracted attention in machine learning in recent years [14, 15]. We recapitulate the method in Appendix B.

In the setting of nonlinear ICA with auxiliary variables, contrastive learning can be exploited by training a classifier to distinguish between a tuple sampled from the joint distribution, which we denote as $(\boldsymbol{x}, \boldsymbol{u})$, and one where $\boldsymbol{u}^*$ is a sample generated from the marginal $p(\boldsymbol{u})$ independently of $\boldsymbol{x}$, $(\boldsymbol{x}, \boldsymbol{u}^*)$. Intuitively, tuples drawn from the former distribution correspond to the same sources $\boldsymbol{s}$, and thus share information, while tuples from the latter correspond to different sources and thus do not share information. Since the marginals of both distributions are equal, the classifier must learn to distinguish between them based on the common information shared by $\boldsymbol{x}$ and $\boldsymbol{u}$; that is, ultimately, $\boldsymbol{s}$.

With this method, the reconstruction of $\boldsymbol{s}$ is only possible up to an invertible scalar "gauge" transformation. This is due to a fundamental ambiguity in the setup of nonlinear ICA and does not represent a limitation of their results;

it can therefore be considered a trivial one. We further comment on this in Appendix A.3.

3 NONLINEAR ICA WITH MULTIPLE VIEWS

We described how naively splitting Equations 1, 2 and 3 into two separate nonlinear ICA problems renders both problems non-identifiable, unless strong assumptions are made on the $\boldsymbol{f}_i$ or the distribution of $\boldsymbol{s}$.

In the Rosetta stone story, awareness that different texts reported on the stele were linked by a common topic helped solving the translation problem; similarly, in our setting, matched observations of the two views are linked through the shared latent variable $\boldsymbol{s}$. Thus the central question we investigate is whether these assumptions can be relaxed by exploiting the structure of the generative model; that is, whether jointly observing $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ provides sufficient constraints to the inverse problem, thus removing the ambiguities present in the vanilla nonlinear ICA setting. We consider a contrastive learning task in which a classifier is trained to distinguish between pairs $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ corresponding to the same $\boldsymbol{s}$ and $(\boldsymbol{x}_1, \boldsymbol{x}_2^*)$ corresponding to different realizations of $\boldsymbol{s}$. As discussed in Section 2, the classifier will be forced to employ the information shared by the simultaneous views in order to distinguish the two classes. As we show, this ultimately results in recovering $\boldsymbol{s}$ (up to unavoidable ambiguities).

For technical reasons discussed in Appendix B, our method requires some stochasticity in the relationship between $\boldsymbol{s}$ and at least one of the $\boldsymbol{x}_i$. However this is not a significant constraint in practice; in most real settings observations are corrupted by noise, and a truly deterministic relationship between $\boldsymbol{s}$ and the $\boldsymbol{x}_i$ would be unrealistic.

We will consider a component-wise independent corruption of our sources, i.e. $\boldsymbol{x}_1 = \boldsymbol{f}_1 \circ \boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1)$ with $g_{1i}(\boldsymbol{s}, \boldsymbol{n}_1) = g_{1i}(s_i, n_{1i})$, where the components of $\boldsymbol{n}_1$ are mutually independent, and similar for $\boldsymbol{x}_2$. The noise variables $\boldsymbol{n}_1, \boldsymbol{n}_2$ and the sources $\boldsymbol{s}$ are assumed to be mutually independent. Note that this only puts constraints on the way the signal is corrupted by the noise, namely $\boldsymbol{g}$, and not on the mixing $\boldsymbol{f}$. We will refer to such $\boldsymbol{g}$ as *component-wise corrupter* throughout, and to its output as *corruption*. In the the vanilla ICA setting, inverting the mixing and recovering the sources $\boldsymbol{s}$ are equivalent; in the setting that we consider, the inversion of the mixing $\boldsymbol{f}$ only implies recovering the sources up to the effect of the corrupter $\boldsymbol{g}$.

We will consider three instances of the general setting, providing identifiability results for each.

I. First we consider the case that only one of the observations, $\boldsymbol{x}_2$, is corrupted with noise. This corresponds, for instance, to a setting in which one accurate measurement device is supplemented with a second noisy device. We show that in this setting it is possible to fully reconstruct $\boldsymbol{s}$ using the noiseless variable (Section 3.1).

II. Next, we consider the case that both variables are corrupted with noise. In this setting, it is possible to recover $\boldsymbol{s}$ up to the corruptions. Furthermore, we show that $\boldsymbol{s}$ can be recovered with arbitrary precision in the limit that the corruptions go to zero (Section 3.2).

III. Finally, we consider the case of having N simultaneous views of the source $\boldsymbol{s}$ rather than just two. When considering the limit $N \to \infty$, we prove sufficient conditions under which it is possible to reconstruct $\boldsymbol{s}$ even if each observation is corrupted by noise (Section 3.3).

To the best of our knowledge, no result of identifiability of latent sources in the case in which only corrupted, mixed versions are observed has been given before.

3.1 ONE NOISELESS VIEW

Consider the generative model

$$\boldsymbol{x}_1 = \boldsymbol{f}_1(\boldsymbol{s}) \tag{5}$$

$$\boldsymbol{x}_2 = \boldsymbol{f}_2(\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n})) \tag{6}$$

$$p(\boldsymbol{s}) = \prod_i p_i(s_i)$$

$$p(\boldsymbol{n}) = \prod_i p_i(n_i) \tag{7}$$

where $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ are invertible, $\boldsymbol{g}$ is a component-wise corrupter, $\boldsymbol{n} \perp\!\!\!\perp \boldsymbol{s}$ and $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are observed. This is represented in Figure 1.

Subject to some assumptions, it is possible to recover $\boldsymbol{s}$ up to the component-wise invertible ambiguity.

Theorem 1. *The difference of the log joint probability and log product of marginals of the observed variables in the generative model specified by Equations 5-7 admits*

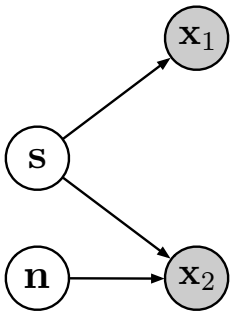

Figure 1: The setting considered in Section 3.1. Two views of the sources are available, one of which, $\boldsymbol{x}_1$, is not corrupted by noise. In this and all other figures, each node is a deterministic function of all its parents in the graph.

the following factorisation:

$$\begin{aligned}
&\log p(\boldsymbol{x}_1, \boldsymbol{x}_2) - \log p(\boldsymbol{x}_1)p(\boldsymbol{x}_2) \\
&= \log p(\boldsymbol{x}_2|\boldsymbol{x}_1) - \log p(\boldsymbol{x}_2) \\
&= \left(\sum_i \alpha_i(s_i, g_i(s_i, n_i)) + \log\det J \right) \\
&\quad - \left(\sum_i \delta_i(g_i(s_i, n_i)) + \log\det J \right) \\
&= \sum_i \alpha_i(s_i, g_i(s_i, n_i)) - \sum_i \delta_i(g_i(s_i, n_i)) \qquad (8)
\end{aligned}$$

where $s_i = f_{1i}^{-1}(\boldsymbol{x}_1)$, $g_i = f_{2i}^{-1}(\boldsymbol{x}_2)$, and J is the Jacobian of the transformation f_2^{-1} (note that the introduced Jacobians cancel). Suppose that

1. *$\boldsymbol{\alpha}$ satisfies the* Sufficiently Distinct Views *assumption (see after this theorem).*
2. *We train a classifier to discriminate between*

$$(\boldsymbol{x}_1, \boldsymbol{x}_2) \text{ vs. } (\boldsymbol{x}_1, \boldsymbol{x}_2^*),$$

   where $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ correspond to the same realization of $\boldsymbol{s}$ and $(\boldsymbol{x}_1, \boldsymbol{x}_2^*)$ correspond to different realizations of $\boldsymbol{s}$.
3. *The classifier is constrained to use a regression function of the form*

$$r(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_i \psi_i(h_i(\boldsymbol{x}_1), \boldsymbol{x}_2)$$

   where $\boldsymbol{h} = (h_1, \ldots, h_n)$ are invertible, smooth and have smooth inverse.

Then, in the limit of infinite data and with universal approximation capacity, $\boldsymbol{h}$ inverts $\boldsymbol{f}_1$ in the sense that the $h_i(\boldsymbol{x}_1)$ recover the independent components of $\boldsymbol{s}$ up to component-wise invertible transformations.

The proof can be found in Appendix D.1. The assumption of invertibility for $\boldsymbol{h}$ could be satisfied by, e.g., the use of normalizing flows [33, 8] or deep invertible networks [24].

We remark that at several points in this paper we consider the difference between two log-probabilities. In all of these cases, the Jacobians introduced by a change of variables cancel out as in Equation 8. For brevity we omit explanation of this fact in the rest of the results.

The *Sufficiently Distinct Views (SDV)* assumption specifies in a technical sense that the two views available are sufficiently different from one another, resulting in more information being available in totality than from each view individually. In the context of Theorem 1, it is an assumption about the log-probability of the *corruption* conditioned on the source. Informally, it demands that the probability distribution of the corruption should vary significantly as a result of conditioning on different values of the source.

Definition 2 (Sufficiently Distinct Views)**.** *Let $\alpha_i(y_i, t_i)$, $i = 1, \ldots, N$ be functions of two arguments. Denote by $\boldsymbol{\alpha}$ the vector of functions and define*

$$\alpha_i'(y_i, t_i) = \partial\alpha_i(y_i, t_i)/\partial t, \qquad (9)$$

$$\alpha_i''(y_i, t_i) = \partial^2\alpha_i(y_i, t_i)/\partial t^2 \qquad (10)$$

$$\boldsymbol{w}_{\boldsymbol{\alpha}}(\boldsymbol{y}, \boldsymbol{t}) = (\alpha_1'', \ldots, \alpha_D'', \alpha_1', \ldots, \alpha_D'). \qquad (11)$$

We say that $\boldsymbol{\alpha}$ satisfies the assumption of Sufficiently Distinct Views (SDV) *if for any value of $\boldsymbol{y}$, there exist $2D$ distinct values $\boldsymbol{t}_j$, $j = 1, \ldots, 2D$ such that the vectors $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_j)$ are linearly independent.*

This is closely related to the Assumption of Variability in [23]. We provide simple cases of conditional log-probability density functions satisfying and violating the SDV assumption in Appendix C.

Theorem 1 shows that by jointly considering the two views, it is possible to recover $\boldsymbol{s}$, in contrast to the single-view setting. This result can be extended to learn the inverse of $\boldsymbol{f}_2$ up to component-wise invertible functions.

Corollary 3. *Consider the setting of Theorem 1, and the alternative factorization of the log joint probability given by*

$$\begin{aligned}
&\log p(\boldsymbol{x}_1, \boldsymbol{x}_2) - \log p(\boldsymbol{x}_1)p(\boldsymbol{x}_2) \\
&= \log p(\boldsymbol{x}_1|\boldsymbol{x}_2) - \log p(\boldsymbol{x}_1) \\
&= \sum_i \gamma_i(s_i, g_i(s_i, n_i)) - \sum_i \beta_i(s_i)\,. \qquad (12)
\end{aligned}$$

Suppose that $\boldsymbol{\gamma}$ satisfies the SDV assumption. Replacing the regression function with

$$r(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_i \psi_i(\boldsymbol{x}_1, h_i(\boldsymbol{x}_2))$$

results in $\boldsymbol{h}$ inverting $\boldsymbol{f}_2$ in the sense that the $h_i(\boldsymbol{x}_2)$ recover the independent components of $\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n})$ up to component-wise invertible transformations.

The proof can be found in Appendix D.2. These two results together mean that it is possible to learn inverses $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ of $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, and therefore to recover $\boldsymbol{s}$ and $\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n})$, up to component-wise intertible functions. Note, however, that doing so requires running two separate algorithms. Furthermore, there is no guarantee that the learned inverses $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are 'aligned' in the sense that for each i the components $\boldsymbol{h}_{1i}(\boldsymbol{x}_1)$ and $\boldsymbol{h}_{2i}(\boldsymbol{x}_2)$ correspond to the same components of $\boldsymbol{s}$.

This problem of misalignment can be resolved by changing the form of the regression function.

Theorem 4. *Consider the settings of Theorem 1 and Corollary 3. Suppose that both $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ satisfy the SDV assumption. Replacing the regression function with*

$$r(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_i \psi_i(h_{1,i}(\boldsymbol{x}_1), h_{2,i}(\boldsymbol{x}_2)) \tag{13}$$

results in $\boldsymbol{h}_1$, $\boldsymbol{h}_2$ inverting $\boldsymbol{f}_1$, $\boldsymbol{f}_2$ in the sense that the $h_{1,i}(\boldsymbol{x}_1)$ and $h_{2,i}(\boldsymbol{x}_2)$ recover the independent components of $\boldsymbol{s}$ and $\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n})$ up to two different component-wise invertible transformations. Furthermore, the two representations are aligned, i.e. for $i \neq j$,

$$h_{1,i}(\mathbf{x}_1) \perp\!\!\!\perp h_{2,j}(\mathbf{x}_2)$$

The proof can be found in Appendix E.

Note that Theorem 4 is *not* a generalisation of Theorem 1 or Corollary 3, since it makes stricter assumptions by imposing the SDV assumption on both $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. In contrast, Theorem 1 and Corollary 3 require that only one is valid for each.

For cases in which finding aligned representations for $\boldsymbol{s}$ and $\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n})$ are desired, Theorem 4 should be applied. If the only goal is recovery of $\boldsymbol{s}$, the assumptions of Theorem 1 are simpler to verify.

In practical applications, the multi-view scenario is useful in multimodal datasets where one of the two acquisition modalities has much higher signal to noise ratio than the other one (e.g., in neuroimaging, when simultaneous fMRI and Optical Imaging recordings are compared). In such cases, jointly exploiting the multiple modalities would help to discern a meaningful and identifiable latent representation which could not be attained through analysis of the more reliable modality alone.

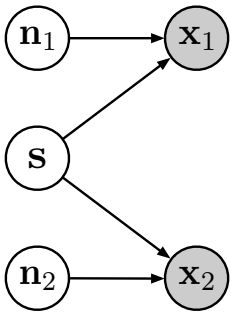

Figure 2: Setting with two views of the sources $\boldsymbol{s}$, both corrupted by noise.

3.1.1 Equivalence with Permutation Contrastive Learning for Time Dependent Sources

Note that the analysis of Theorem 1 covers the case of temporally dependent stationary sources analyzed in [20]. Indeed, if it is further assumed that $\boldsymbol{s}$ and $\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n})$ are uniformly dependent [20, Definition 1], they can be seen as a pair of subsequent time points of an ergodic stationary stochastic process for which the analysis of Theorem 1 of [20] would hold. In other words, we can define a stochastic process as $p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t) := p(\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n})|\boldsymbol{s})$. Note that while the two formulations are theoretically equivalent, our view offers a wider applicability as it covers the asynchronous sensing of $\boldsymbol{s}$, provided that multiple measurements (i.e. $\boldsymbol{x}_1, \boldsymbol{x}_2$) are available; additionally, our *Sufficiently Distinct Views* assumption does not necessarily imply uniform dependency. Furthermore, while [20] considers a generative model of the form $\boldsymbol{x}(t) = \boldsymbol{f}(\boldsymbol{s}(t))$, thus constraining the mixing function to be the same for any two data points $\boldsymbol{x}(t_1)$, $\boldsymbol{x}(t_2)$, in our setting we consider two different mixing functions, $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, for the two different views. Finally, we study this setting as an intermediate step for the following two sections, in which no deterministic function of the sources is observed, learning to invert any of the $\boldsymbol{f}_i$ can only recover $\boldsymbol{s}$ up to the corruption operated by $\boldsymbol{g}$.

3.2 TWO NOISY VIEWS

We next consider the setting in which both variables are corrupted by noise. Consider the following generative model (represented in Figure 2):

$$\begin{aligned} \boldsymbol{x}_1 &= \boldsymbol{f}_1(\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1)) \\ \boldsymbol{x}_2 &= \boldsymbol{f}_2(\boldsymbol{g}_2(\boldsymbol{s}, \boldsymbol{n}_2)) \,, \end{aligned}$$

where all variables take value in $\mathbb{R}^D$, and $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ are nonlinear, invertible, deterministic functions, $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ are component-wise corrupters, and $\boldsymbol{s}$ and the $\boldsymbol{n}_i$ are independent with independent components. This class of

models generalizes the setting of Section 3.1 since by taking $\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1) = \boldsymbol{s}$ we reduce to the case of one noiseless observation.

The difference $\log p(\boldsymbol{x}_1, \boldsymbol{x}_2) - \log p(\boldsymbol{x}_1)p(\boldsymbol{x}_2)$ admits similar factorizations to those given in Equations 8 and 12:

$$\begin{aligned}
&\log p(\boldsymbol{x}_1, \boldsymbol{x}_2) - \log p(\boldsymbol{x}_1)p(\boldsymbol{x}_2) \\
&= \log p(\boldsymbol{x}_1|\boldsymbol{x}_2) - \log p(\boldsymbol{x}_1) \\
&= \sum_i \eta_i(g_{1i}(s_i, n_{i1}), g_{2i}(s_i, n_{2i})) - \sum_i \theta_i(g_{1i}(s_i, n_{1i}) \qquad (14) \\
&= \log p(\boldsymbol{x}_2|\boldsymbol{x}_1) - \log p(\boldsymbol{x}_2) \\
&= \sum_i \lambda_i(g_{2i}(s_i, n_{2i}), g_{1i}(s_i, n_{1i})) - \sum_i \mu_i(g_{2i}(s_i, n_{2i})) \qquad (15)
\end{aligned}$$

Since we only have access to corrupted observations, exact recovery of $\boldsymbol{s}$ is not possible. Nonetheless, a generalization of Theorem 4 holds showing that the $\boldsymbol{f}_i$ can be inverted and $\boldsymbol{s}$ recovered up to the corruptions induced by the $\boldsymbol{n}_i$ via $\boldsymbol{g}_i$.

Theorem 5. *Suppose that $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$ satisfy the SDV assumption. The algorithm described in Theorem 1 with regression function specified in Equation 13 results in $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ inverting $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ in the sense that the $h_{1,i}(\boldsymbol{x}_1)$ and $h_{2,i}(\boldsymbol{x}_2)$ recover the independent components of $\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1)$ and $\boldsymbol{g}_2(\boldsymbol{s}, \boldsymbol{n}_2)$ up to two different component-wise invertible transformations. Furthermore, the two representations are aligned, i.e. for $i \neq j$,*

$$h_{1,i}(\mathbf{x}_1) \perp\!\!\!\perp h_{2,j}(\mathbf{x}_2)$$

The proof can be found in Appendix E.

We can thus recover the common source $\boldsymbol{s}$ up to the corruptions $\boldsymbol{g}_i(\boldsymbol{s}, \boldsymbol{n}_i)$. In the limit of the magnitude of one of the noise variables going to zero, the reconstruction of the sources $\boldsymbol{s}$ attained through the corresponding view is exact up to the component-wise invertible functions, as stated in the following corollary.

Corollary 6. *Let $\boldsymbol{n}_1^{(k)} = \frac{1}{k} \cdot \tilde{\boldsymbol{n}}$ for $k \in \mathbb{N}$, where $\tilde{\boldsymbol{n}} \in \mathbb{R}^D$ is a fixed random variable, and $\boldsymbol{n}_2$ be a random variable that does not depend on k. Let $\boldsymbol{h}_1^{(k)}, \boldsymbol{h}_2^{(k)}$ be the output of the algorithm specified by Theorem 5 with noise variables $\boldsymbol{n}_1^{(k)}$ and $\boldsymbol{n}_2$.*

Suppose that the corrupters $\boldsymbol{g}_i$ satisfy the following two criteria:

i) $\exists \boldsymbol{a} \in \mathbb{R}^D_{>0}$ *s.t.* $\left|\frac{\partial \boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n})}{\partial \boldsymbol{n}}\right|_{\boldsymbol{n}=0} \leq \boldsymbol{a}$ *for all* $\boldsymbol{s}$

ii) $\exists \boldsymbol{b} \in \mathbb{R}^D_{>0}$ *s.t.* $0 < \frac{\partial \boldsymbol{g}_1(\boldsymbol{s}, 0)}{\partial \boldsymbol{s}} \leq \boldsymbol{b}$

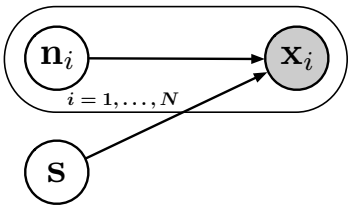

Figure 3: Setting with N corrupted views of the sources.

Then, denoting by $\boldsymbol{E}$ the set of all scalar, invertible functions, we have that

$$\lim_{k \to \infty} \inf_{\boldsymbol{e} \in \boldsymbol{E}} \left\| \boldsymbol{s} - \boldsymbol{e}(\boldsymbol{h}_1^{(k)}(\boldsymbol{x}_1)) \right\| = 0$$

The proof can be found in Appendix F.

Corollary 6 implies that in the limit of small noise, the sources $\boldsymbol{s}$ can be recovered exactly. Condition *i)* upper bounds the influence of $\boldsymbol{n}$ on the corruption: we can not hope to retrieve $\boldsymbol{s}$ if $\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n})$ contains too little signal. Condition *ii)* ensures that the function $\boldsymbol{g}$ is invertible with respect to $\boldsymbol{s}$ when $\boldsymbol{n}$ is equal to zero. If this were not satisfied, some information about $\boldsymbol{s}$ would be washed out by $\boldsymbol{g}$ even in absence of noise. This would make recovery of $\boldsymbol{s}$ trivially impossible.

3.3 MULTIPLE NOISY VIEWS

The results of Section 3.2 state that in the two noisy view setting, $\boldsymbol{s}$ can be recovered up to the corruptions. In the limit that the magnitude of the noises goes to zero, the uncorrupted $\boldsymbol{s}$ can be recovered. The intuition is that the less noise there is, the more information each observation provides about $\boldsymbol{s}$.

In this section we consider the multi-view setting, where N distinct noisy views of $\boldsymbol{s}$ are available,

$$\boldsymbol{x}_i = \boldsymbol{f}_i(\boldsymbol{g}_i(\boldsymbol{s}, \boldsymbol{n}_i)) \ \ , i = 1, \ldots, N \, ,$$

and the noise variables $\boldsymbol{n}_i$ are mutually independent, as represented in Figure 3. Since each view provides additional information about $\boldsymbol{s}$, we ask: in the limit as $N \to \infty$, is it possible to reconstruct $\boldsymbol{s}$ exactly?

By applying Theorem 5 to the pair $(\boldsymbol{x}_1, \boldsymbol{x}_i)$ it is possible to recover $(\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1), \boldsymbol{g}_i(\boldsymbol{s}, \boldsymbol{n}_i))$ such that the components are aligned, but up to different component-wise invertible functions $\boldsymbol{k}_1$ and $\boldsymbol{k}_i$. Running the algorithm on a different pair $(\boldsymbol{x}_1, \boldsymbol{x}_j)$ will result in recovery up to different component-wise invertible functions $\boldsymbol{k}'_1$ and $\boldsymbol{k}'_j$.

Note that these will *not* necessarily result in $\boldsymbol{k}_i \circ \boldsymbol{g}_i(\boldsymbol{s}, \boldsymbol{n}_i)$ and $\boldsymbol{k}'_j \circ \boldsymbol{g}_j(\boldsymbol{s}, \boldsymbol{n}_j)$ being aligned with each other. However, the components of $\boldsymbol{k}_1 \circ \boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1)$ and $\boldsymbol{k}'_1 \circ \boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1)$

are the same, up to permutation and component-wise invertible functions. This permutation can therefore be undone by performing independence testing between each pair of components. Components that are 'different' will be independent; those that are the same will be deterministically related. Therefore, they can be used as a reference to permute the components of $\boldsymbol{k}'_j$ and make it aligned with $\boldsymbol{k}_i$.

The problem is then how to combine the information from each aligned $\boldsymbol{k}_i \circ \boldsymbol{g}_i(\boldsymbol{s}, \boldsymbol{n}_i)$ to more precisely identify $\boldsymbol{s}$. The fact that the components are recovered up to *different* scalar invertible functions makes combining information from different views non-trivial.

As a first step in this direction, we consider the special case that each $\boldsymbol{g}_i$ acts additively and each $\boldsymbol{n}_i$ is zero mean and each of $\boldsymbol{s}$ and the $\boldsymbol{n}_i$ are independent with independent components.

$$\left.\begin{array}{l} \boldsymbol{x}_i = \boldsymbol{f}_i(\boldsymbol{s} + \boldsymbol{n}_i) \\ \mathbb{E}\boldsymbol{n}_i = 0 \end{array}\right\} \quad i \in \mathbb{N} \tag{16}$$

Suppose to begin with that we are able to recover each $\boldsymbol{s}+\boldsymbol{n}_i$ *without* the usual component-wise invertible functions. Then, writing $\boldsymbol{n}$ to denote all of the $\boldsymbol{n}_i$, it is possible to estimate $\boldsymbol{s}$ as

$$\boldsymbol{s} \approx \Omega^N(\boldsymbol{s}, \boldsymbol{n}) = \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{s} + \boldsymbol{n}_i)\,.$$

Subject to mild conditions on the rate of growth of the variances $\mathrm{Var}(\boldsymbol{n}_i)$ as $i \to \infty$, Kolmogorov's strong law implies that $\Omega^N(\boldsymbol{s}, \boldsymbol{n})$ is a good approximation to $\boldsymbol{s}$ as $N \to \infty$ in the sense that $\Omega^N(\boldsymbol{s}, \boldsymbol{n}) \xrightarrow{a.s.} \boldsymbol{s}$. This implies moreover that it is possible to reconstruct the $\boldsymbol{n}_i$ by considering the residue $R_i^N(\boldsymbol{s}, \boldsymbol{n}) = (\boldsymbol{s} + \boldsymbol{n}_i) - \Omega^N(\boldsymbol{s}, \boldsymbol{n}) \xrightarrow{a.s.} \boldsymbol{n}_i$.

In the presence of the unknown functions $\boldsymbol{k}_i$, we would be able to reconstruct $\boldsymbol{s}$ and the $\boldsymbol{n}_i$ if we were able to identify the inverses $\boldsymbol{e}_i = \boldsymbol{k}_i^{-1}$ for each i. For any component-wise invertible functions $\boldsymbol{e}_i$, define

$$\Omega_{\mathbf{e}}^N(\boldsymbol{s}, \boldsymbol{n}) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{e}_i \circ \boldsymbol{k}_i(\boldsymbol{s} + \boldsymbol{n}_i)$$
$$R_{\boldsymbol{e},i}^N(\boldsymbol{s}, \boldsymbol{n}) = \boldsymbol{e}_i \circ \boldsymbol{k}_i(\boldsymbol{s} + \boldsymbol{n}_i) - \Omega_{\boldsymbol{e}}^N(\boldsymbol{s}, \boldsymbol{n}).$$

$\boldsymbol{e}_i$ is something we can choose and $\boldsymbol{k}_i(\boldsymbol{s} + \boldsymbol{n}_i) = \boldsymbol{h}_i(\boldsymbol{x}_i)$ is the output of the algorithm, and hence $\Omega_{\mathbf{e}}^N(\boldsymbol{s}, \boldsymbol{n})$ and $R_{\boldsymbol{e},i}^N(\boldsymbol{s}, \boldsymbol{n})$ are random variables with known distributions. Subject to mild conditions, the dependence of these quantities on most or all of the $\boldsymbol{n}_i$ becomes increasingly small as N grows and disappears in the limit $N \to \infty$.

Lemma 7. *Suppose that the sequence* $\mathbb{E}_{\boldsymbol{n}}[\Omega_{\mathbf{e}}^N(\boldsymbol{s}, \boldsymbol{n})] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{n}_i}[\boldsymbol{e}_i \circ \boldsymbol{k}_i(\boldsymbol{s} + \boldsymbol{n}_i)]$ *converges as* $N \to \infty$ *for almost all* $\boldsymbol{s}$*, and write*

$$\Omega_{\mathbf{e}}(\boldsymbol{s}) = \lim_{N\to\infty} \mathbb{E}_{\boldsymbol{n}}[\Omega_{\mathbf{e}}^N(\boldsymbol{s}, \boldsymbol{n})].$$

Suppose further that there exists K *such that* $V_{\mathbf{e}_i} = \mathrm{Var}\left(\boldsymbol{e}_i \circ \boldsymbol{g}_i(\boldsymbol{s} + \boldsymbol{n}_i)\right) \leq K$ *for all* i*. Then*

$$\Omega_{\mathbf{e}}^N(\boldsymbol{s}, \boldsymbol{n}) \xrightarrow{a.s.} \Omega_{\mathbf{e}}(\boldsymbol{s})$$
$$R_{\mathbf{e},i}^N(\boldsymbol{s}, \boldsymbol{n}) \xrightarrow{a.s.} R_{\mathbf{e},i}(\boldsymbol{s}, \boldsymbol{n}_i) = \boldsymbol{e}_i \circ \boldsymbol{k}_i(\boldsymbol{s} + \boldsymbol{n}_i) - \Omega_{\mathbf{e}}(\boldsymbol{s})$$

The proof can be found in Appendix G. Given some choice of $\boldsymbol{e}$, we can think of $\Omega_{\mathbf{e}}(\boldsymbol{s})$ and $R_{\mathbf{e},i}(\boldsymbol{s}, \boldsymbol{n}_i)$ as our putative candidates for $\boldsymbol{s}$ and $\boldsymbol{n}_i$ respectively. As discussed earlier, if we could identify $\boldsymbol{e}_i = \boldsymbol{k}_i^{-1}$, then we would have $\Omega_{\mathbf{e}}(\boldsymbol{s}) = \boldsymbol{s}$ and $R_{\mathbf{e},i}(\boldsymbol{s}, \boldsymbol{n}_i) = \boldsymbol{n}_i$, and thus $\Omega_{\mathbf{e}}$ and $R_{\mathbf{e},i}$ would satisfy the same independences and other statistical properties as $\boldsymbol{s}$ and $\boldsymbol{n}_i$ respectively. Can we use these properties as criteria to identify good choices of $\boldsymbol{e}_i$?

The following theorem gives a set of sufficient conditions under which each $\boldsymbol{e}_i$ inverts $\boldsymbol{k}_i$ up to some affine ambiguity which is the same for every i.

Theorem 8. *Suppose there exists* $C > 0$ *such that* $\mathit{Var}(\boldsymbol{n}_i) \leq C$ *for all* i *and let* $\mathcal{G}_K = \left\{\{\boldsymbol{e}_i\} \text{ s.t.}\right.$

$$V_{\boldsymbol{e}_i} \leq K \;\forall i \tag{17}$$
$$\Omega_{\boldsymbol{e}}(\boldsymbol{s}) < \infty \text{ for almost all } \boldsymbol{s} \tag{18}$$
$$R_{\boldsymbol{e},i} \perp\!\!\!\perp R_{\boldsymbol{e},j}\; \forall i \neq j, \tag{19}$$
$$\mathbb{E}R_{\boldsymbol{e},i} = 0\; \forall i \tag{20}$$
$$\left.R_{\boldsymbol{e},i}(\boldsymbol{s}, \boldsymbol{n}_i) = R_{\boldsymbol{e},i}(\boldsymbol{n}_i)\; \forall i\right\} \tag{21}$$

Then,

$$\mathcal{G}_K \subseteq \left\{\{\boldsymbol{\alpha}\boldsymbol{k}_i^{-1} + \boldsymbol{\beta}\} \;:\; \boldsymbol{\alpha} \in \mathbb{R}^D_{\neq 0},\, \boldsymbol{\beta} \in \mathbb{R}^D\right\}$$

where $\boldsymbol{\alpha}\boldsymbol{k}_i^{-1}$ *denotes the element-wise product with the scalar elements of* $\boldsymbol{\alpha}$*. If* $K \geq \mathit{Var}(\boldsymbol{s}) + C$*, then* $\{\boldsymbol{k}_i^{-1}\} \in \mathcal{G}_K$*, and so* $\mathcal{G}_K$ *is non-empty for* K *sufficiently large.*

The proof can be found in Appendix H. It follows that it is possible recover $\boldsymbol{s}$ and $\boldsymbol{n}_i$ up to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ via $\Omega_{\mathbf{e}}(\boldsymbol{s}) = \boldsymbol{\alpha}\boldsymbol{s} + \boldsymbol{\beta}$ and $R_{\boldsymbol{e},i}(\boldsymbol{n}_i) = \boldsymbol{\alpha}\boldsymbol{n}_i$.

We remark that each of the conditions 17–20 can be verified from known information. We conjecture that condition 21 can be relaxed to assuming the verifiable condition of independence between $\Omega_{\mathbf{e}}(\boldsymbol{s})$ and $R_{\mathbf{e},i}(\boldsymbol{s}, \boldsymbol{n}_i)$ for all i along with additional regularity assumptions on the functional form of $R_{\mathbf{e},i}$ (e.g. smoothness).

To conclude, Theorem 8 provides sufficient conditions under which it is possible to fully reconstruct $\boldsymbol{s}$ with

corrupted views. In contrast to previous results in Sections 3.1 and 3.2, this result leverages infinitely many corrupted views rather than vanishingly small corruption of finitely many views.

4 RELATED WORK

A central concept in our work is that of multiple simultaneous views and joint extraction of features from them. We briefly review some related work considering similar settings.

4.1 CANONICAL CORRELATION ANALYSIS

Given two (or more) random variables, the goal of Canonical Correlation Analysis (CCA) [18] is to find a corresponding pair of linear subspaces that have high cross-correlation, so that each component within one of the subspaces is correlated with a single component from the other subspace [5]. In dealing with correlation instead of independence, CCA is more closely related to Principal Component Analysis (PCA) than to ICA.

CCA can be interpretated probabilistically [4] and is equivalent to maximum likelihood estimation in a graphical model which is a special case of that depicted in Figure 2. The differences compared to our setting are (i) the latent components retrieved in CCA are forced to be uncorrelated, whereas our method is retrieves independent components; (ii) in CCA, mappings between the sources $\boldsymbol{s}$ and $\boldsymbol{x}$ are linear, whereas our method allows for nonlinear mappings.

At a high level, the model we consider in Section 3.2 is to CCA as nonlinear ICA is to PCA. Nonlinear extensions of the basic CCA framework have been proposed [26, 13, 3, 29], but identifiability results in the sense we consider in this paper are lacking.

4.2 MULTI-VIEW LATENT VARIABLE MODELS

Bearing a strong resemblance to our considered setting, [27] proposes a sequence of diffusion maps to find the common source of variability captured by multiple sensors, discarding irrelevant sensor-specific effects. It computes the distance among the samples measured by different sensors to form a similarity matrix for the measurements of each sensor; each similarity matrix is then associated to a diffusion operator, which is a Markov matrix by construction. A Markov chain is then run by alternately applying these Markov matrices on the initial state. During these Markovian dynamics, sensor specific information will eventually vanish, and the final state will only contain information on the common source. While the method focuses on recovering the common information in the form of a parametrization of the common variable, our method both inverts the mixing mechanisms of each view and recovers the common latent variables.

[39] proves identifiability for multi-view, latent variable models, unifying previously proposed spectral techniques [2]. However, while the setting is similar to the one considered in this work, both the objectives and the employed methods are different. The paper considers the setting in which L variables X_l, $l = 1, \ldots, L$ are observed; additionally, there exists an unobserved latent variable H, such that conditional distributions $P(X_l|H)$ are independent. While the setting bears obvious similarities with our multi-view ICA, the method proposed in [39] is aimed at learning the mixture parameters, rather than the exact realization of latent variables. Their method is based on the mean embedding of distributions in a Reproducing Kernel Hilbert Space and a result of identifiability for the parameters of the mean embeddings of $P(H)$ and $P(X|H)$ is proved.

Another related field of study is multi-view clustering, which considers a multiview setting and aims at performing clustering on a given dataset, see e.g. [11] and [25]. While related to our setting, this line of work is different from it in two key ways. Firstly, clustering can be thought of as assigning a discrete latent label per datapoint. In contrast, our setting seeks to recover a continuous latent vector per datapoint. Second, since no underlying generative model with discrete latent variable is assumed, identifiability results are not given.

4.3 HALF-SIBLING REGRESSION

Half-sibling regression [35] is a method to reconstruct a source from noisy observations by exploiting other sources that are affected by the same noise process but otherwise independent from it.

Suppose that a latent variable of interest Q is not directly available, and that we can only observe corrupted versions of it, denoted as Y, where the corruption is due to a noise N. Without knowledge of N, it is impossible to reconstruct Q. However, if one or more additional variables X, also influenced by N, are observed, we can exploit them to model the effect of N on Y by regressing Y on X.

Subtracting this from the observed Y recovers the latent variable Q up to a constant offset, provided that (1) the additivity assumption

$$Y = Q + f(N)$$

holds, and (2) that Y contains sufficient information

about $f(N)$. Analogous to our aim of recovering $\boldsymbol{s}$, the goal of half-sibling regression is not to infer only the distribution of Q, but rather the random variable itself (almost surely).

5 DISCUSSION AND CONCLUSION

We presented identifiability results in a novel setting by extending the formalism of nonlinear ICA. We have investigated different scenarios of multi-view latent variable models and provided theoretical proofs on the possibility of inverting the mixing function and recovering the sources in each case. Our results thus extend the scarce literature on identifiability for nonlinear ICA models.

In the classical noiseless ICA setting, the deterministic relationship between the sources and observations means that inverting the mixing function and recovering the sources are equivalent. In contrast, we consider views of corrupted versions of the common sources, resulting in the decoupling of the demixing and retrieval of the sources. Remarkably, Theorem 8 points towards the possibility of simultaneously solving the two problems in the limit of infinitely many views.

Classical nonlinear ICA is provably non-identifiable because a single view is not sufficiently informative to resolve non-trivial ambiguities when recovering the sources. While many papers in the ICA literature have explored placing restrictions either on the source distribution or on the form of the mixing to resolve these ambiguities, in this paper we consider exploiting additional views to constrain the inverse problem. Clearly, if a second view is identical to the first, then nothing is gained by its observation. Hence, in order for the second view to assist in resolving ambiguity, it must be sufficiently different from the first. This is the intuition behind the technical assumption of *sufficiently distinct views*.

Typically, noise is a nuisance variable that would be preferably non-existent. In our setting, however, the noise variables acting on the sources are a crucial component, without which the contrastive learning approach could not be applied. Furthermore, the assumption of sufficiently distinct views is ultimately an assumption about the complexity of the joint distribution of the (corrupted) sources corresponding to each view. Without the noise variables the sufficiently distinct views assumption could not hold.

Our setting is relevant in a number of practical real-world applications, namely in all datasets that include multiple distinct measurements of related phenomena. In practice, it may be better to think of the noise variables rather as intrinsic sources of variability specific to each view. In most practical applications this would probably not be a significant limitation due to the prevalence of stochasticity in real-world systems.

An exemplary application of our method can be found in the field of neuroimaging. Consider a study involving a cohort of subjects (perceivers), measuring their response to the presentation of the same stimulus. One of the key problems in the field is how to extract a shared response from all subjects despite high inter-subject variability and complex nonlinear mappings between latent source and observation [7, 16]. Our results provide principled ways to extract and decompose the components of the shared response. In particular, the setting described in our model is suited to account for the high variability of the responses throughout the cohort, since the measurement corresponding to each subject is given by a combination of individual variability and shared response.

Looking to the future, we note that Theorem 8 builds on the setting of Theorem 5 which only makes use of pairwise information from the observations. A natural extension of this work should investigate algorithms that explicitly make use of $N > 2$ views, which we conjecture would allow relaxation of the additivity assumption on the corruptions. Furthermore, Theorem 8 provides results that only hold for the asymptotic limit as the number of views becomes large. Other extensions to this result could include analysis of the case of finitely many views.

Acknowledgements

Thanks to Krikamol Muandet for providing his office for fruitful discussions, to Matthias Bauer and Manuel Wüthrich for proofreading and to Lucia Busso for interesting input about linguistics.

References

[1] Naomi Allen, Cathie Sudlow, Paul Downey, Tim Peakman, John Danesh, Paul Elliott, John Gallacher, Jane Green, Paul Matthews, Jill Pell, et al. Uk biobank: Current status and what it means for epidemiology. *Health Policy and Technology*, 1(3):123–126, 2012.

[2] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[3] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255, 2013.

[4] Francis R Bach and Michael I Jordan. A probabilistic interpretation of canonical correlation analysis. 2005.

[5] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

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

[7] Po-Hsuan Cameron Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James Haxby, and Peter J Ramadge. A reduced-dimension fMRI shared response model. In *Advances in Neural Information Processing Systems*, pages 460–468, 2015.

[8] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, pages 6572–6583, 2018.

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

[10] 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.

[11] Virginia R De Sa. Spectral clustering with two views. In *ICML workshop on learning with multiple views*, pages 20–27, 2005.

[12] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[13] Kenji Fukumizu, Francis R Bach, and Arthur Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(Feb):361–383, 2007.

[14] 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.

[15] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

[16] James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke, and Peter J Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011.

[17] Timo Honkela, Aapo Hyvärinen, and Jaakko J Väyrynen. WordICA-emergence of linguistic representations for words by independent component analysis. *Natural Language Engineering*, 16(3):277–308, 2010.

[18] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer, 1992.

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

[20] Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of Temporally Dependent Stationary Sources. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 460–469, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

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

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

[23] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 859–868. PMLR, 16–18 Apr 2019.

[24] Jörn-Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon. i-RevNet: Deep Invertible Networks. In *ICLR 2018 - International Conference on Learning Representations*, Vancouver, Canada, April 2018.

[25] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. In *Advances in neural information processing systems*, pages 1413–1421, 2011.

[26] Pei Ling Lai and Colin Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(05):365–377, 2000.

[27] Roy R Lederman and Ronen Talmon. Learning the geometry of common latent variables using alternating-diffusion. *Applied and Computational Harmonic Analysis*, 44(3):509–536, 2018.

[28] Martin J McKeown and Terrence J Sejnowski. Independent component analysis of fMRI data: examining the assumptions. *Human brain mapping*, 6(5-6):368–372, 1998.

[29] Tomer Michaeli, Weiran Wang, and Karen Livescu. Nonparametric canonical correlation analysis. In *International Conference on Machine Learning*, pages 1967–1976, 2016.

[30] Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper LR Andersson, et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523, 2016.

[31] Danielle Nuzillard and Albert Bijaoui. Blind source separation and analysis of multispectral astronomical images. *Astronomy and Astrophysics Supplement Series*, 147(1):129–138, 2000.

[32] Erkki Oja, Kimmo Kiviluoto, and Simona Malaroiu. Independent component analysis for financial time series. In *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373)*, pages 111–116. IEEE, 2000.

[33] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

[34] Hiroshi Sawada, Ryo Mukai, and Shoji Makino. Direction of arrival estimation for multiple source signals using independent component analysis. In *Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings.*, volume 2, pages 411–414. IEEE, 2003.

[35] Bernhard Schölkopf, David W Hogg, Dun Wang, Daniel Foreman-Mackey, Dominik Janzing, Carl-Johann Simon-Gabriel, and Jonas Peters. Modeling confounding by half-sibling regression. *Proceedings of the National Academy of Sciences*, 113(27):7391–7398, 2016.

[36] Meredith A Shafto, Lorraine K Tyler, Marie Dixon, Jason R Taylor, James B Rowe, Rhodri Cusack, Andrew J Calder, William D Marslen-Wilson, John Duncan, Tim Dalgleish, et al. The cambridge centre for ageing and neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC neurology*, 14(1):204, 2014.

[37] Amit Singer and Ronald R Coifman. Non-linear independent component analysis with diffusion maps. *Applied and Computational Harmonic Analysis*, 25(2):226–239, 2008.

[38] Viktor Pavlovich Skitovich. Linear forms of independent random variables and the normal distribution law. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, 18(2):185–200, 1954.

[39] Le Song, Animashree Anandkumar, Bo Dai, and Bo Xie. Nonparametric estimation of multi-view latent variable models. In *International Conference on Machine Learning*, pages 640–648, 2014.

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

[41] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The WU-Minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.