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

$$x_1 = f_1(s) \tag{1}$$
$$x_2 = f_2(s) \tag{2}$$
$$p(s) = \prod_i p_i(s_i), \tag{3}$$

---

[*]Equal contribution.

where $x_1, x_2, s \in \mathbb{R}^D$ and f_1, f_2 are arbitrary smooth and invertible transformations of the latent variable $s = (s_1, \ldots, s_D)$ with mutually independent components. The goal is to recover s, undoing the mixing induced by the f_i, in the case where only observations of x_1 and 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 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 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

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 $x = f(s)$ are available, where 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 u and a change in the independence properties. If the *unconditional independence* is substituted with a *conditional independence* given the auxiliary variable u, i.e.

$$\log p(s|u) = \sum_i q_i(s_i, u),\qquad(4)$$

for some functions q_i, the model becomes identifiable. The conditional independence statement in Equation 4 can be interpreted as positing that u is a parent of the sources s. A further assumption on the effect of variations in u on x, called *variability* in the paper, is required. Intuitively, it demands that u has a sufficiently diverse influence on 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 (x, u), and one where u^* is a sample generated from the marginal $p(u)$ independently of x, (x, u^*). Intuitively, tuples drawn from the former distribution correspond to the same sources 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 x and u; that is, ultimately, s.

With this method, the reconstruction of 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 f_i or the distribution of 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 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 x_1 and 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 (x_1, x_2) corresponding to the same s and (x_1, x_2^*) corresponding to different realizations of 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 s (up to unavoidable ambiguities).

For technical reasons discussed in Appendix B, our method requires some stochasticity in the relationship between s and at least one of the 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 s and the x_i would be unrealistic. We will consider a component-wise independent corruption of our sources, i.e. $x_1 = f_1 \circ g_1(s, n_1)$ with $g_{1i}(s, n_1) = g_{1i}(s_i, n_{1i})$, where the components of n_1 are mutually independent, and similar for x_2. The noise variables n_1, n_2 and the sources s are assumed to be mutually independent. Note that this only puts constraints on the way the signal is corrupted by the noise, namely g, and not on the mixing f. We will refer to such 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 s are equivalent; in the setting that we consider, the inversion of the mixing f only implies recovering the sources up to the effect of the corrupter 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, 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 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 s up to the corruptions. Furthermore, we show that 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 s rather than just two. When considering the limit $N \to \infty$, we prove sufficient conditions under which it is possible to reconstruct 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

$$x_1 = f_1(s) \tag{5}$$
$$x_2 = f_2(g(s, n)) \tag{6}$$
$$p(s) = \prod_i p_i(s_i)$$
$$p(n) = \prod_i p_i(n_i) \tag{7}$$

where f_1 and f_2 are invertible, g is a component-wise corrupter, $n \perp\!\!\!\perp s$ and x_1 and x_2 are observed. This is represented in Figure 1.

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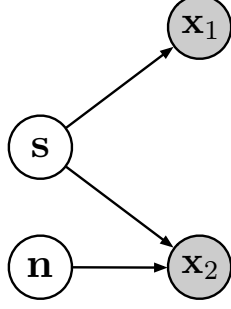
Figure 1: The setting considered in Section 3.1. Two views of the sources are available, one of which, $\boldsymbol{x}_1$, is not corrupted by noise. In this and all other figures, each node is a deterministic function of all its parents in the graph.

the following factorisation:

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

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*

$$\alpha_i'(y_i, t_i) = \partial \alpha_i(y_i, t_i)/\partial t, \quad (9)$$
$$\alpha_i''(y_i, t_i) = \partial^2 \alpha_i(y_i, t_i)/\partial t^2 \quad (10)$$
$$\boldsymbol{w_\alpha}(\boldsymbol{y}, \boldsymbol{t}) = (\alpha_1'', \ldots, \alpha_D'', \alpha_1', \ldots, \alpha_D'). \quad (11)$$

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*

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

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 $h_{1i}(\boldsymbol{x}_1)$ and $h_{2i}(\boldsymbol{x}_2)$ 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*

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

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.
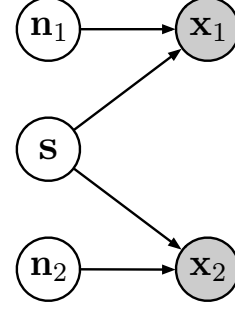

Figure 2: Setting with two views of the sources $\boldsymbol{s}$, both corrupted by noise.

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], 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):

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

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 $g_1(s, n_1) = s$ we reduce to the case of one noise-less observation.

The difference $\log p(x_1, x_2) - \log p(x_1)p(x_2)$ admits similar factorizations to those given in Equations 8 and 12:

$$\log p(x_1, x_2) - \log p(x_1)p(x_2)$$
$$= \log p(x_1|x_2) - \log p(x_1)$$
$$= \sum_i \eta_i(g_{1i}(s_i, n_{i1}), g_{2i}(s_i, n_{2i})) - \sum_i \theta_i(g_{1i}(s_i, n_{1i}))$$
$$\tag{14}$$

$$= \log p(x_2|x_1) - \log p(x_2)$$
$$= \sum_i \lambda_i(g_{2i}(s_i, n_{2i}), g_{1i}(s_i, n_{1i})) - \sum_i \mu_i(g_{2i}(s_i, n_{2i}))$$
$$\tag{15}$$

Since we only have access to corrupted observations, exact recovery of s is not possible. Nonetheless, a generalization of Theorem 4 holds showing that the f_i can be inverted and s recovered up to the corruptions induced by the n_i via g_i.

Theorem 5. *Suppose that η and λ satisfy the SDV assumption. The algorithm described in Theorem 1 with regression function specified in Equation 13 results in h_1 and h_2 inverting f_1 and f_2 in the sense that the $h_{1,i}(x_1)$ and $h_{2,i}(x_2)$ recover the independent components of $g_1(s, n_1)$ and $g_2(s, n_2)$ 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 s up to the corruptions $g_i(s, n_i)$. In the limit of the magnitude of one of the noise variables going to zero, the reconstruction of the sources s attained through the corresponding view is exact up to the component-wise invertible functions, as stated in the following corollary.

Corollary 6. *Let $n_1^{(k)} = \frac{1}{k} \cdot \tilde{n}$ for $k \in \mathbb{N}$, where $\tilde{n} \in \mathbb{R}^D$ is a fixed random variable, and n_2 be a random variable that does not depend on k. Let $h_1^{(k)}, h_2^{(k)}$ be the output of the algorithm specified by Theorem 5 with noise variables $n_1^{(k)}$ and n_2.*

Suppose that the corrupters g_i satisfy the following two criteria:

i) $\exists a \in \mathbb{R}_{>0}^D$ *s.t.* $\left| \frac{\partial g_1(s,n)}{\partial n} \right|_{n=0} \leq a$ *for all s*

ii) $\exists b \in \mathbb{R}_{>0}^D$ *s.t.* $0 < \frac{\partial g_1(s,0)}{\partial s} \leq b$
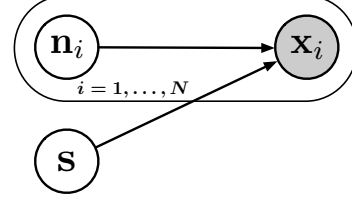

Figure 3: Setting with N corrupted views of the sources.

Then, denoting by E the set of all scalar, invertible functions, we have that

$$\lim_{k \to \infty} \inf_{e \in E} \left\| s - e(h_1^{(k)}(x_1)) \right\| = 0$$

The proof can be found in Appendix F.

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

3.3 MULTIPLE NOISY VIEWS

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

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

$$x_i = f_i(g_i(s, n_i)) \;\; , i = 1, \ldots, N \,,$$

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

By applying Theorem 5 to the pair (x_1, x_i) it is possible to recover $(g_1(s, n_1), g_i(s, n_i))$ such that the components are aligned, but up to different component-wise invertible functions k_1 and k_i. Running the algorithm on a different pair (x_1, x_j) will result in recovery up to different component-wise invertible functions k'_1 and k'_j.

Note that these will *not* necessarily result in $k_i \circ g_i(s, n_i)$ and $k'_j \circ g_j(s, n_j)$ being aligned with each other. However, the components of $k_1 \circ g_1(s, n_1)$ and $k'_1 \circ g_1(s, n_1)$

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 k_j' and make it aligned with k_i.

The problem is then how to combine the information from each aligned $k_i \circ g_i(s, n_i)$ to more precisely identify 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 g_i acts additively and each n_i is zero mean and each of s and the n_i are independent with independent components.

$$\left. \begin{array}{l} x_i = f_i(s + n_i) \\ \mathbb{E}n_i = 0 \end{array} \right\} \quad i \in \mathbb{N} \qquad (16)$$

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

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

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

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

$$\Omega_e^N(s, n) = \frac{1}{N} \sum_{i=1}^{N} e_i \circ k_i(s + n_i)$$

$$R_{e,i}^N(s, n) = e_i \circ k_i(s + n_i) - \Omega_e^N(s, n).$$

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

Lemma 7. *Suppose that the sequence* $\mathbb{E}_n[\Omega_e^N(s, n)] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{n_i}[e_i \circ k_i(s + n_i)]$ *converges as* $N \to \infty$ *for almost all* s, *and write*

$$\Omega_e(s) = \lim_{N \to \infty} \mathbb{E}_n[\Omega_e^N(s, n)].$$

Suppose further that there exists K such that $V_{e_i} = \mathrm{Var}(e_i \circ g_i(s + n_i)) \leq K$ for all i. Then

$$\Omega_e^N(s, n) \xrightarrow{a.s.} \Omega_e(s)$$
$$R_{e,i}^N(s, n) \xrightarrow{a.s.} R_{e,i}(s, n_i) = e_i \circ k_i(s + n_i) - \Omega_e(s)$$

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

The following theorem gives a set of sufficient conditions under which each e_i inverts k_i up to some affine ambiguity which is the same for every i.

Theorem 8. *Suppose there exists $C > 0$ such that $\mathrm{Var}(n_i) \leq C$ for all i and let $\mathcal{G}_K = \{\{e_i\}$ s.t.*

$$V_{e_i} \leq K \; \forall i \qquad (17)$$
$$\Omega_e(s) < \infty \; \text{for almost all } s \qquad (18)$$
$$R_{e,i} \perp\!\!\!\perp R_{e,j} \; \forall i \neq j, \qquad (19)$$
$$\qquad (20)$$
$$\mathbb{E}R_{e,i} = 0 \; \forall i \qquad (21)$$
$$R_{e,i}(s, n_i) = R_{e,i}(n_i) \; \forall i \} \qquad (22)$$

Then,

$$\mathcal{G}_K \subseteq \{\{\alpha k_i^{-1} + \beta\} \; : \; \alpha \in \mathbb{R}_{\neq 0}^D, \; \beta \in \mathbb{R}^D\}$$

where αk_i^{-1} denotes the element-wise product with the scalar elements of α. If $K \geq \mathrm{Var}(s) + C$, then $\{k_i^{-1}\} \in \mathcal{G}_K$, 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 s and n_i up to α and β via $\Omega_e(s) = \alpha s + \beta$ and $R_{e,i}(n_i) = \alpha n_i$.

We remark that each of the conditions 17–21 can be verified from known information. We conjecture that condition 22 can be relaxed to assuming the verifiable condition of independence between $\Omega_e(s)$ and $R_{e,i}(s, n_i)$ for all i along with additional regularity assumptions on the functional form of $R_{e,i}$ (e.g. smoothness).

To conclude, Theorem 8 provides sufficient conditions under which it is possible to fully reconstruct 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 s and 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 multi-view 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 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.

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 analy-

sis. 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 Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR.

[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.

APPENDIX

A ON THE UNIDENTIFIABILITY OF NONLINEAR ICA

The purpose of this section is to briefly review the proof of unidentifiablity of nonlinear ICA as [22]: In this section we assume the most general conventional form of nonlinear ICA where the generative model follows:

$$\mathbf{x} = \mathbf{f}(\boldsymbol{s}) \tag{23}$$

where $\boldsymbol{s}$ are the independent sources and $\mathbf{x}$ are mixed signals. In the following, we show how to construct a function $\mathbf{g} : R^n \to R^n$ so that the components $\mathbf{y} = \mathbf{g}(\mathbf{x})$ are independent. More importantly, we show that this construction is by no means unique.

A.1 EXISTENCE

The proposed method in [22] is a generalization of the famous Gram-Schmidt orthogonalization. Given m independent variables, $y_1, \ldots, y_m$ and a variable x, one constructs a new variable $y_{m+1} = g(y_1, \ldots, y_m, x)$ so that the set $y_1 s, \ldots, y_{m+1}$ is mutually independent. The construction process is defined recursively as follows. Assume we have m independent random variables $y_1, \ldots, y_m$ with uniform distribution in $[0, 1]^m$. x is any random variable and $a_1, \ldots, a_m, b$ are some nonrandom scalars. Next, we define

$$g\left(a_1, \ldots, a_m, b; p_{y,x}\right) = p(x \le b | y_1 = a_1, \ldots, y_m = a_m)$$
$$= \frac{\int_{-\infty}^{b} p_{y,x}\left(a_1, \ldots, a_m, \xi\right) \mathrm{d}\xi}{p_y\left(a_1, \ldots, a_m\right)} \tag{24}$$

Theorem 1 of [22] says that the random variable defined as $y_{m+1} = g(y_1, \ldots, y_m, x)$ is independent from the $y_1 \ldots, y_m$ and $y_1, \ldots, y_{m+1}$ are uniformly distributed in the unit cube $[0, 1]^{m+1}$.

A.2 NON-UNIQUENESS

In the previous section, it was shown that there exists a mapping $\mathbf{g}$ that transforms any random vector $\mathbf{x}$ into a uniformly distributed random vector $\mathbf{y} = \mathbf{g}(\mathbf{x})$. Here, we show that the construction of $\mathbf{g}$ is not unique and this non-Uniqueness can be caused by several factors.

- A linear transformation $\mathbf{x}'$ can precede the nonlinear map $\mathbf{f}$ and then compute the independent components $\mathbf{y}' = \mathbf{g}'(\mathbf{x}')$ where $\mathbf{g}'$ is computed as describe in the previous section. The new map $\mathbf{g}'$ gives
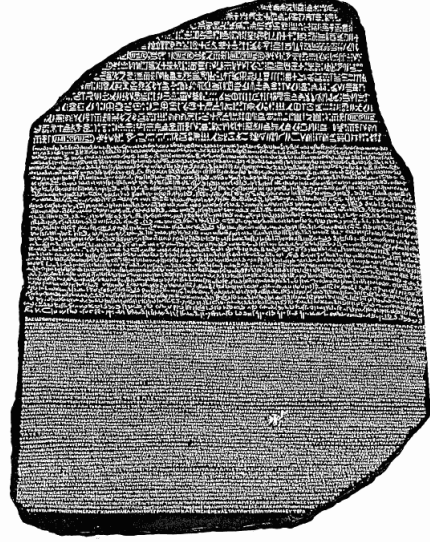

Figure 4: The Rosetta Stone, a stele found in 1799, inscribed with three versions of a decree issued at Memphis, Egypt in 196 BC. The top and middle texts are in Ancient Egyptian using hieroglyphic script and Demotic script, respectively, while the bottom is in Ancient Greek. (Source: Wikipedia)

a new decomposition of $\mathbf{x}$ into independent components $\mathbf{y}'$ which can not be trivially reduced to $\mathbf{y}$.

- An element-wise function $\mathbf{h}$ can apply on the independent sources $\boldsymbol{s}$ first to give new sources $\boldsymbol{s}'$ such that $s_i' = h_i(s_i)$. Constructing the solution g for these new scaled version of sources gives a new decomposition into independent components.

- Assume a class of measure-preserving automorphisms $\mathbf{h} : [0, 1]^n \to [0, 1]^n$. The mapping $\mathbf{h}$ does not change the probability distribution of a uniformly distributed random variable in n-dimensional hypercube. The composition $\mathbf{h} \circ \mathbf{g}$ gives another solution to nonlinear ICA. Therefore, the class of measure-preserving automorphisms gives a parameterization of the solutions to nonlinear ICA introducing a class of non-trivial indeterminacies.

If only independence among the components matters, it is possible to construct a mapping $\boldsymbol{y} = G(\boldsymbol{x})$ such that y_i is independent of y_j for $i \ne j$ and uniformely distributed in $[0, 1]^n$. This shows that at least one solution exists. The non-uniqueness of the solution can be shown by parameterising a class of infinitely many solutions. Once $\boldsymbol{y}$ is found with above conditions, any measure-preserving automorphism $f : [0, 1]^n \to [0, 1]^n$ can be used to parameterize G as $f \circ G$, suggesting that there

are infinitely many solutions to nonlinear ICA whose relations are nontrivial.

A.3 THE SCALAR INVERTIBLE FUNCTION GAUGE

Another indeterminacy is element-wise functions f_i applying on y_i which suggests another dimension of ambiguity. Non-Gaussianity cannot help here since we can construct any marginal distribution by combining the CDF of the observed variable with the inverse CDF of the target marginal distribution. This indeterminacy is in some sense unavoidable and is related to the fact that in linear ICA recovery of the sources is possible up to a scalar multiplicative ambiguity.

B WHY DOES CLASSIFICATION RESULT IN THE LOG RATIO?

Let us suppose that a variable X is drawn with equal probability from two distributions P_0 and P_1 with densities $p_0(x)$ and $p_1(x)$ respectively. We train a classifier $D : x \mapsto [0, 1]$ to estimate the posterior probability that a particular realization of X was drawn from P_0 with the cross entropy loss, i.e. the parameters of D are chosen to minimize

$$L(D) = \mathbb{E}_{X \sim P_0}\left[-\log D(X)\right] + \mathbb{E}_{X \sim P_1}\left[-\log(1 - D(X))\right].$$

As shown in, for instance, [14], the global optimum of this loss occurs when $D(x) = \frac{p_0(x)}{p_0(x) + p_1(x)}$, which can be rewritten as

$$D(x) = \frac{1}{1 + p_1(x)/p_0(x)} \tag{25}$$

$$= \frac{1}{1 + \exp(-\log(p_0(x)/p_1(x)))} \tag{26}$$

$$\tag{27}$$

Recall that in our setting, the function $r(x_1, x_2)$ is trained to classify between the two cases that (x_1, x_2) is drawn from the joint distribution $\mathbb{P}_{x_1, x_2}$ (*class 0*) or the product of marginals $\mathbb{P}_{x_1}\mathbb{P}_{x_2}$ (*class 1*). $r(x_1, x_2)$ is trained so that $\frac{1}{1 + \exp(-r(x_1, x_2))}$ estimates the posterior probability of (x_1, x_2) belonging to class 0. By comparing to Equation 26, it can be seen that

$$r(x_1, x_2) = \log\left(p(x_1, x_2)/p(x_1)p(x_2)\right)$$
$$= \log p(x_1|x_2) - \log p(x_1)$$
$$= \log p(x_2|x_1) - \log p(x_2)$$

Note that in order for the classification trick of contrastive learning to be useful, the variables x_1 and x_2 cannot be deterministically related. If this is the case, the log-ratio is everywhere either 0 or ∞ and hence the learned features are not useful.

To see why this is the case, suppose that x_1, and x_2 are each N-dimensional vectors. If they are deterministically related, $p(x_1, x_2)$ puts mass on an N-dimensional submanifold of a $2N$-dimensional space. On the other hand, $p(x_1)p(x_2)$ will put mass on a $2N$-dim manifold since it is the product of two distributions each of which are N-dimensional.

In this case, the distributions $p(x_1, x_2)$ and $p(x_1)p(x_2)$ are therefore not absolutely continuous with respect to one another and thus the log-ratio is ill-defined: $p(x_1, x_2)/p(x_1)p(x_2) = \infty$ at any point (x_1, x_2) at which $p(x_1, x_2)$ puts mass and zero at points where $p(x_1)p(x_2)$ puts mass and $p(x_1, x_2)$ does not.

C THE SUFFICIENTLY DISTINCT VIEWS ASSUMPTION

We give the following two examples to provide intuition about the Sufficiently Distinct Views (SDV) assumption - one regarding a case in which it does not hold, and another one in which it does.

A simple case in which the assumption does not hold is when the conditional probability of z given s is Gaussian, as in

$$p(z|s) = \frac{1}{Z} \exp\left[-\sum_i (z_i - s_i)^2/(2\sigma_i^2)\right], \tag{28}$$

where Z is the normalization factor, $Z = (2\pi)^{n/2}\prod_i \sigma_i$. Since taking second derivatives of the log-probability with respect to s_i results in constants, it can be easily shown that there is no way to find $2D$ vectors z_j, $j = 1, \ldots, 2D$, such that the corresponding $w(s, z_j)$ (see Definition 1) are linearly independent.

The fact that the assumption breaks down in this case is reminiscent of the breakdown in the case of Gaussianity for linear ICA. Interestingly, in our work, the true latent

sources **are** allowed to be Gaussian. In fact, the distribution of s does not enter the expression above.

An example in which the SDV assumption does hold is a conditional pdf given by

$$p(\boldsymbol{z}|\boldsymbol{s}) = \frac{1}{Z(\boldsymbol{s})} \exp \left[-\sum_i (z_i^2 s_i^2 + z_i^4 s_i^4) \right] , \quad (29)$$

where $Z(\boldsymbol{s})$ is again a normalization function. Proving that this distribution satisfies the SDV assumption requires a few lines of computation. The idea is that $\boldsymbol{w}(\boldsymbol{s}, \boldsymbol{z})$ can be written as the product of a matrix and vector which are functions only of $\boldsymbol{s}$ and $\boldsymbol{z}$ respectively. Once written in this form, it is straightforward to show that the columns of the matrix are linearly independent for almost all values of $\boldsymbol{s}$ and that $2D$ linearly independent vectors can be realized by different choices of $\boldsymbol{z}$.

D   PROOF OF THEOREM 1 AND COROLLARY 3

D.1   PROOF OF THEOREM 1

This proof is mainly inspired by the techniques employed by [23].

Proof. We have to show that, upon convergence, $h_i(\boldsymbol{x}_1)$ are s.t.

$$h_i(\boldsymbol{x}_1) \perp\!\!\!\perp h_j(\boldsymbol{x}_1), \forall i \neq j$$

We start by writing the difference in log-densities of the two classes:

$$\sum_i \psi_i(h_i(\boldsymbol{x}_1), \boldsymbol{x}_2) = \sum_i \alpha_i(\boldsymbol{f}_{1,i}^{-1}(\boldsymbol{x}_1), \boldsymbol{f}_{2,i}^{-1}(\boldsymbol{x}_2)) +$$
$$- \sum_i \delta_i(\boldsymbol{f}_{2,i}^{-1}(\boldsymbol{x}_2))$$

We now make the change of variables

$$\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x}_1)$$
$$\boldsymbol{v}(\boldsymbol{y}) = \boldsymbol{f}_1^{-1}(\boldsymbol{h}^{-1}(\boldsymbol{y}))$$
$$\boldsymbol{t} = \boldsymbol{f}_2^{-1}(\boldsymbol{x}_2))$$

and rewrite the first equation in the following form:

$$\sum_i \psi_i(y_i, \boldsymbol{x}_2) = \sum_i \alpha_i(v_i(\boldsymbol{y}), t_i) \quad (30)$$
$$- \sum_i \delta_i(t_i) \quad (31)$$

We take derivatives with respect to y_j, $y_{j'}$, $j \neq j'$, of the LHS and RHS of equation 39. Adopting the conventions in 9 and 10 and

$$v_i^j(\boldsymbol{y}) = \partial v_i(\boldsymbol{y})/\partial y_j \quad (32)$$
$$v_i^{jj'}(\boldsymbol{y}) = \partial^2 v_i(\boldsymbol{y})/\partial y_j \partial y_{j'} , \quad (33)$$

we have

$$\sum_i \alpha_i''(v_i(\boldsymbol{y}), t_i) v_i^j(\boldsymbol{y}) v_i^{j'}(\boldsymbol{y})$$
$$+ \alpha_i'(v_i(\boldsymbol{y}), t_i) v_i^{jj'}(\boldsymbol{y}) = 0 ,$$

where taking derivative w.r.t. y_j and y_j' for $j \neq j'$ makes LHS equal to zero, since the LHS has functions which depend only one y_i each. If we now rearrange our variables by defining vectors $\boldsymbol{a}_i(\boldsymbol{y})$ collecting all entries $v_i^j(\boldsymbol{y}) v_i^{j'}(\boldsymbol{y})$, $j = 1, \ldots, n$, $j' = 1, \ldots, j - 1$, and vectors $\boldsymbol{b}_i(\boldsymbol{y})$ with the variables $v_i^j(\boldsymbol{y}) v_i^{j'}(\boldsymbol{y})$, $j = 1, \ldots, n$, $j' = 1, \ldots, j - 1$, the above equality can be rewritten as

$$\sum_i \alpha_i''(v_i(\boldsymbol{y}), t_i) \boldsymbol{a}_i(\boldsymbol{y})$$
$$+ \alpha_i'(v_i(\boldsymbol{y}), t_i)) \boldsymbol{b}_i(\boldsymbol{y}) = 0 .$$

The above expression can be recast in matrix form,

$$\boldsymbol{M}(\boldsymbol{y}) \boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}) = 0 ,$$

where $\boldsymbol{M}(\boldsymbol{y}) = (\boldsymbol{a}_1(\boldsymbol{y}), \ldots, \boldsymbol{a}_n(\boldsymbol{y}), \boldsymbol{b}_1(\boldsymbol{y}), \ldots, \boldsymbol{b}_n(\boldsymbol{y}))$ and $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}) = (\alpha_1'', \ldots, \alpha_n'', \alpha_1', \ldots, \alpha_n')$. $\boldsymbol{M}(\boldsymbol{y})$ is therefore a $n(n-1)/2 \times 2n$ matrix, and $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t})$ is a $2n$ dimensional vector.

To show that $\boldsymbol{M}(\boldsymbol{y})$ is equal to zero, we invoke the SDV assumption. This implies the existence of $2n$ linearly independent $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_j)$. It follows that

$$\boldsymbol{M}(\boldsymbol{y})[\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_1), \ldots, \boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_{2n})] = 0 ,$$

and hence $\boldsymbol{M}(\boldsymbol{y})$ is zero by elementary linear algebraic results. It follows that $v_i^j(\boldsymbol{y}) \neq 0$ for at most one value of j, since otherwise the product of two non-zero terms would appear in one of the entries of $\boldsymbol{M}(\boldsymbol{y})$, thus rendering it non-zero. Thus v_i is a function only of one y_j.

Observe that $\boldsymbol{v}(\boldsymbol{y}) = \boldsymbol{s}$. We have just proven that $v_i(y_{\pi(i)}) = s_i$. Since v_i is invertible, it follows that $h_{\pi(i)}(\boldsymbol{x}_1) = y_{\pi(i)} = v_i^{-1}(s_i)$ and hence the components of $\boldsymbol{h}(\boldsymbol{x}_1)$ recover the components of $\boldsymbol{s}$ up to the invertible component-wise ambiguity given by $\boldsymbol{v}$, and the permutation ambiguity.

$\square$

D.2 PROOF OF COROLLARY 3

Proof. This follows exactly by repeating the proof of Theorem 1 where the roles of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are exchanged and the regression function in the statement of the corollary is used. $\square$

E PROOF OF THEOREMS 4 AND 5

Theorem 4 is a special case of Theorem 5 by considering the case $\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1) = \boldsymbol{s}$. We therefore prove only the more general Theorem 5.

Proof. We have to show that, upon convergence, $h_i(\boldsymbol{x}_1)$ and $k_i(\boldsymbol{x}_2)$ are such that

$$h_{1,i}(\boldsymbol{x}_1) \perp\!\!\!\perp h_{1,j}(\boldsymbol{x}_1), \forall i \neq j \qquad (34)$$

$$h_{2,i}(\boldsymbol{x}_2) \perp\!\!\!\perp h_{2,j}(\boldsymbol{x}_2), \forall i \neq j \qquad (35)$$

$$h_{1,i}(\boldsymbol{x}_1) \perp\!\!\!\perp h_{2,j}(\boldsymbol{x}_2), \forall i \neq j. \qquad (36)$$

We start by exploiting Equations 14 and 15 to write the difference in log-densities of the two classes

$$\sum_i \psi_i(h_{1,i}(\boldsymbol{x}_1), h_{2,i}(\boldsymbol{x}_2))$$
$$= \sum_i \eta_i(\boldsymbol{f}_{1,i}^{-1}(\boldsymbol{x}_1), \boldsymbol{f}_{2,i}^{-1}(\boldsymbol{x}_2)) - \sum_i \theta_i(\boldsymbol{f}_{1,i}^{-1}(\boldsymbol{x}_1))$$
$$\qquad (37)$$
$$= \sum_i \lambda_i(\boldsymbol{f}_{2,i}^{-1}(\boldsymbol{x}_2), \boldsymbol{f}_{1,i}^{-1}(\boldsymbol{x}_1)) - \sum_i \mu_i(\boldsymbol{f}_{2,i}^{-1}(\boldsymbol{x}_2))$$
$$\qquad (38)$$

We now make the change of variables

$$\boldsymbol{y} = \boldsymbol{h}_1(\boldsymbol{x}_1)$$
$$\boldsymbol{t} = \boldsymbol{h}_2(\boldsymbol{x}_2)$$
$$\boldsymbol{v}(\boldsymbol{y}) = \boldsymbol{f}_1^{-1}(\boldsymbol{h}_1^{-1}(\boldsymbol{y}))$$
$$\boldsymbol{u}(\boldsymbol{t}) = \boldsymbol{f}_2^{-1}(\boldsymbol{h}_2^{-1}(\boldsymbol{t}))$$

and rewrite equation 37 in the following form:

$$\sum_i \psi_i(y_i, t_i)$$
$$= \sum_i \eta_i(v_i(\boldsymbol{y}), u_i(\boldsymbol{t})) - \sum_i \theta_i(v_i(\boldsymbol{y})) \qquad (39)$$

We first want to prove the condition in Equation 34. We will show this is true by proving that

$$v_i(\boldsymbol{y}) \equiv v_i(y_{\pi(i)}) \qquad (40)$$

for some permutation of the indices π with respect to the indexing of the sources $\boldsymbol{s} = (s_1, \ldots, s_D)$.

We take derivatives with respect to y_j, $y_{j'}$, $j \neq j'$, of the LHS and RHS of equation 39, yielding

$$\sum_i \eta_i''(v_i(\boldsymbol{y}), u_i(\boldsymbol{t})) v_i^j(\boldsymbol{y}) v_i^{j'}(\boldsymbol{y})$$
$$+ \sum_i \eta_i'(v_i(\boldsymbol{y}), u_i(\boldsymbol{t})) v^{jj'}(\boldsymbol{y}) = 0$$

If we now rearrange our variables by defining vectors $\boldsymbol{a}_i(\boldsymbol{y})$ collecting all entries $v_i^j(\boldsymbol{y}) v_i^{j'}(\boldsymbol{y})$, $j = 1, \ldots, n$, $j' = 1, \ldots, j-1$, and vectors $\boldsymbol{b}_i(\boldsymbol{y})$ with the variables $v_i^j(\boldsymbol{y}) v_i^{j'}(\boldsymbol{y})$, $j = 1, \ldots, n$, $j' = 1, \ldots, j-1$, the above equality can be rewritten as

$$\sum_i \eta_i''(v_i(\boldsymbol{y}), u_i(\boldsymbol{t})) \boldsymbol{a}_i(\boldsymbol{y})$$
$$+ \eta_i'(v_i(\boldsymbol{y}), u_i(\boldsymbol{t})) \boldsymbol{b}_i(\boldsymbol{y}) = 0.$$

Again following [23], we recast the above formula in matrix form,

$$\boldsymbol{M}(\boldsymbol{y}) \boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}) = 0, \qquad (41)$$

where $\boldsymbol{M}(\boldsymbol{y}) = (\boldsymbol{a}_1(\boldsymbol{y}), \ldots, \boldsymbol{a}_n(\boldsymbol{y}), \boldsymbol{b}_1(\boldsymbol{y}), \ldots, \boldsymbol{b}_n(\boldsymbol{y}))$ and $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}) = (\eta_1'', \ldots, \eta_n'', \eta_1', \ldots, \eta_n')$. $\boldsymbol{M}(\boldsymbol{y})$ is therefore a $n(n-1)/2 \times 2n$ matrix, and $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t})$ is a $2n$ dimensional vector.

To show that $\boldsymbol{M}(\boldsymbol{y})$ is equal to zero, we invoke the SDV assumption on $\boldsymbol{\eta}$. This implies the existence of $2n$ linearly independent $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_j)$. It follows that

$$\boldsymbol{M}(\boldsymbol{y})[\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_1), \ldots, \boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_{2n})] = 0,$$

and hence $\boldsymbol{M}(\boldsymbol{y})$ is zero by elementary linear algebraic results. It follows that $v_i^j(\boldsymbol{y}) \neq 0$ for at most one value of j, since otherwise the product of two non-zero terms would appear in one of the entries of $\boldsymbol{M}(\boldsymbol{y})$, thus rendering it non-zero. Thus v_i is a function only of one $y_j = y_{\pi(i)}$.

Observe that $\boldsymbol{v}(\boldsymbol{y}) = \boldsymbol{s}$. We have just proven that $v_i(y_{\pi(i)}) = s_i$. Since v_i is invertible, it follows that $h_{\pi(i)}(\boldsymbol{x}_1) = y_{\pi(i)} = v_i^{-1}(s_i)$ and hence the components of $\boldsymbol{h}(\boldsymbol{x}_1)$ recover the components of $\boldsymbol{s}$ up to the invertible component-wise ambiguity given by $\boldsymbol{v}$, and the permutation ambiguity.

For the condition in Equation 35, we need

$$u_i(\boldsymbol{t}) \equiv u_i(t_{\tilde{\pi}(i)}), \qquad (42)$$

where the permutation $\tilde{\pi}$ doesn't need to be equal to π. By symmetry, exactly the same argument as used to

prove the condition in Equation 40 holds, by replacing (v, y, η, θ) with (u, t, λ, μ), noting that the SDV assumption is also assumed for λ.

We have shown that $y = h_1(x_1)$ and $t = h_2(x_2)$ estimate $g_1(s, n_1)$ and $g_2(s, n_2)$ up to two different gauges of all possible scalar invertible functions.

A remaining ambiguity could be that the two representations might be misaligned; that is, defining $z_1 = g_1(s, n_1)$ and $z_2 = g_2(s, n_2)$, while

$$z_{1,i} \perp\!\!\!\perp z_{2,j} \forall i \neq j \tag{43}$$

we might have

$$y_{\pi(i)} \perp\!\!\!\perp t_{\tilde{\pi}(j)} \forall i \neq j,$$

where $\pi(i)$, $\tilde{\pi}(i)$ are two different permutations of the indices $i = 1, \ldots, n$. We want to show that this ambiguity is also resolved; that means, our goal is to show that

$$y_i \perp\!\!\!\perp t_j, \quad \forall i \neq j \tag{44}$$

We recall that, by definition, we have $v_i(y_{\pi(i)}) = z_{1,i}$ and $u_j(t_{\tilde{\pi}(j)}) = z_{2,j}$. Then, due to equation 43,

$$v_i(y_{\pi(i)}) \perp\!\!\!\perp u_j(t_{\tilde{\pi}(j)}) \quad \forall i \neq j \tag{45}$$
$$\implies y_{\pi(i)} \perp\!\!\!\perp t_{\tilde{\pi}(j)} \quad \forall i \neq j \tag{46}$$
$$\implies y_i \perp\!\!\!\perp t_{\tilde{\pi} \circ \pi^{-1}(j)} \quad \forall i \neq j, \tag{47}$$

where the implication 45-46 follows from invertibility of v_i and u_j, and the implication 46-47 follows from considering that, given that we know 46, we can define $l = \pi(j)$ and $k = \pi(i)$ and have

$$y_k \perp\!\!\!\perp t_{\tilde{\pi} \circ \pi^{-1}(l)} \quad \forall k \neq l.$$

Define

$$\tau = \tilde{\pi} \circ \pi^{-1}$$

and note that it is a permutation. Then

$$y_i \perp\!\!\!\perp t_{\tau(j)} \forall i \neq j \tag{48}$$

Fix any particular i. Our goal is to show that for any $j \neq i$ the independence relation in Equation 44 holds. There are two possibilities:

i $\tau(i) = i$

ii $\tau(i) \neq i$

In the first case, τ restricted to the set $\{1, \ldots, D\} \setminus \{i\}$ is still a permutation, and thus considering the independences of Equation 48 for all $j \neq i$ implies each of the independences of Equation 44 and we are done.

Let us consider the second case. Then,

$$\exists l \in \{1, \ldots, D\} \setminus \{i\} \text{ s.t. } l = \tau(i).$$

We then need to prove

$$y_i \perp\!\!\!\perp t_l, \tag{49}$$

which is the only independence implied by Equation 44 which is not implied by Equation 48.

In order to do so, we rewrite equation 39, yielding

$$\sum_m \psi_m(y_m, t_m)$$
$$= \sum_m \eta_m(v_m(y_{\pi(m)}), u_m(t_{\tilde{\pi}(m)})) - \sum_m \theta_i(v_m(y_{\pi(m)})) \tag{50}$$

We now take derivative with respect to y_i and t_l in 49; noting that $\tilde{\pi}^{-1}(l) = \pi^{-1}(i)$, we get

$$0 = \frac{\partial^2}{\partial v_{\pi^{-1}(i)} \partial u_{\pi^{-1}(i)}} \eta_{\pi^{-1}(i)}(v_{\pi^{-1}(i)}(y_i), u_{\pi^{-1}(i)}(t_l))$$
$$\times \frac{\partial}{\partial y_i} v_{\pi^{-1}(i)}(y_i) \frac{\partial}{\partial t_l} u_{\pi^{-1}(i)}(t_l) \tag{51}$$

Since $v_{\pi^{-1}(i)}(y_i)$ is a smooth and invertible function of its argument, the set of y_i such that $\frac{\partial}{\partial y_i} v_{\pi^{-1}(i)}(y_i) = 0$ has measure zero. Similarly, $\frac{\partial}{\partial t_l} u_{\pi^{-1}(i)}(t_l) = 0$ on a set of measure zero.

It therefore follows that

$$\frac{\partial}{\partial y_i} v_{\pi^{-1}(i)}(y_i) \frac{\partial}{\partial t_l} u_{\pi^{-1}(i)}(t_l) \neq 0$$

almost everywhere and hence that

$$\frac{\partial^2}{\partial v_{\pi^{-1}(i)} \partial u_{\pi^{-1}(i)}} \eta_{\pi^{-1}(i)}(v_{\pi^{-1}(i)}(y_i), u_{\pi^{-1}(i)}(t_l)) = 0. \tag{52}$$

almost everywhere. We can thus conclude that

$$\eta_{\pi^{-1}(i)}(v_{\pi^{-1}(i)}(y_i), u_{\pi^{-1}(i)}(t_l)) =$$
$$\eta^y_{\pi^{-1}(i)}(v_{\pi^{-1}(i)}(y_i)) + \eta^t_{\pi^{-1}(i)}(u_{\pi^{-1}(i)}(t_l))$$

This in turn implies that, for some functions A and B, we can write

$$\log p(z_{1,\pi^{-1}(i)} | z_{2,\pi^{-1}(i)}) - \log p(z_{1,\pi^{-1}(i)})$$
$$= A(v_{\pi^{-1}(i)}(y_i)) + B(u_{\pi^{-1}(i)}(t_l))$$

and therefore

$$\log p(z_{1,\pi^{-1}(i)}, z_{2,\pi^{-1}(i)}) = C(v_{\pi^{-1}(i)}(y_i)) + D(u_{\pi^{-1}(i)}(t_l))$$

for some functions C and D. This decomposition of the log-pdf implies

$$z_{1,\pi^{-1}(i)} \perp\!\!\!\perp z_{2,\pi^{-1}(i)}$$
$$\implies z_{1,\pi^{-1}(i)} \perp\!\!\!\perp z_{2,\tilde{\pi}^{-1}(l)}$$
$$\implies v_{\pi^{-1}(i)}(y_i) \perp\!\!\!\perp u_{\tilde{\pi}^{-1}(l)}(t_l)$$
$$\implies y_i \perp\!\!\!\perp t_l,$$

where the last implication holds due to invertibility of $v_{\pi^{-1}(i)}$ and $u_{\tilde{\pi}^{-1}(l)}$.

We have thus concluded the proof.

$\square$

F PROOF OF COROLLARY 6

Proof. Denoting by $d_1^{(k)}$ the component-wise invertible ambiguity up to which $g(s, n_1^{(k)})$ is recovered, we have that

$$\inf_{e \in E} \mathbb{E}_{x_1} \left[\left\| s - e(h_1^{(k)}(x_1)) \right\|_2^2 \right] \tag{53}$$

$$= \inf_{e \in E} \mathbb{E}_{(n_1^{(k)}, s)} \left[\left\| s - e \circ d_1^{(k)} \circ g_1(s, n_1^{(k)}) \right\|_2^2 \right] \tag{54}$$

$$= \inf_{\tilde{e} \in E} \mathbb{E}_{(n_1^{(k)}, s)} \left[\left\| s - \tilde{e} \circ g_1(s, n_1^{(k)}) \right\|_2^2 \right] \tag{55}$$

$$\leq \mathbb{E}_{(n_1^{(k)}, s)} \left[\left\| s - e^* \circ g_1(s, n_1^{(k)}) \right\|_2^2 \right] \tag{56}$$

The lower bound holds for any $e^* \in E$ by definition of infimum and in particular for $e^* = g_1|_{n=0}^{-1}$, the existence of which is guaranteed by the assumptions on g_1. Taking a Taylor expansion of $e^* \circ g_1(s, n_1^{(k)})$ around $n_1^{(k)} = 0$ yields

$$\mathbb{E}_{(n_1^{(k)}, s)} \left[\left\| s - e^* \circ g_1(s, 0) \right.\right.$$

$$\left.\left. + \frac{\partial e^*}{\partial g_1} \frac{\partial g_1(s, 0)}{\partial n_1^{(k)}} \cdot n_1^{(k)} + \mathcal{O}(\|n_1^{(k)}\|^2) \right\|_2^2 \right]$$

$$= \mathbb{E}_{(n_1^{(k)}, s)} \left[\left\| \frac{\partial e^*}{\partial g_1} \frac{\partial g_1(s, 0)}{\partial n_1^{(k)}} \cdot n_1^{(k)} + \mathcal{O}(\|n_1^{(k)}\|^2) \right\|_2^2 \right]$$

$$\longrightarrow 0 \text{ as } k \longrightarrow \infty$$

where the last equality follows from fact that $e^* = g|_{n=0}^{-1}$ and the convergence follows from the fact that $n_1^{(k)} \longrightarrow 0$ as $k \to \infty$. $\square$

G PROOF OF LEMMA 7

We will make crucial use of *Kolmogorov's strong law*:

Theorem 9. *Suppose that X_n is a sequence of independent (but not necessarily identically distributed) random variables with*

$$\sum_{n=1}^{\infty} \frac{1}{n^2} \text{Var}[X_n] < \infty$$

Then,

$$\frac{1}{N} \sum_{n=1}^{N} X_n - \mathbb{E}[X_n] \xrightarrow{a.s.} 0$$

Fix s and consider $\Omega_e^N(s, n)$ as a random variable with randomness induced by n. We will show that for almost all s this converges n-almost surely to a constant, and hence $\Omega_e^N(s, n)$ converges almost surely to a function of s.

The law of total expectation says that

$$\text{Var}_{s, n_i}[e_i \circ k_i(s + n_i)]$$
$$= \mathbb{E}_s[V_i(s)] + \text{Var}_s[\mathbb{E}_{n_i}[e_i \circ k_i(s + n_i)]]$$
$$\geq \mathbb{E}_s[V_i(s)].$$

Since by assumption $\text{Var}_{s, n_i}[e_i \circ k_i(s + n_i)] \leq K$, we have that

$$\mathbb{E}_s \left[\sum_{i=1}^{\infty} \frac{V_i(s)}{i^2} \right] \leq \frac{K\pi^2}{6}$$

and therefore $\sum_{i=1}^{\infty} \frac{V_i(s)}{i^2} < \infty$ with probability 1 over s, else the expectation above would be unbounded since $V_i(s) \geq 0$.

We have further that for almost all s,

$$\Omega_e(s) = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} E_{e_i}(s)$$

exists. Therefore, for almost all s the conditions of Kolmogorov's strong law are met by $\Omega_e^N(s, n)$ and so

$$\Omega_e^N(s, n) - \mathbb{E}_n[\Omega_e^N(s, n)] \xrightarrow{n-a.s.} 0$$

Since $\mathbb{E}_n[\Omega_e^N(s, n)] \xrightarrow{n-a.s.} \Omega_e(s)$, it follows that

$$\Omega_e^N(s, n) \xrightarrow{n-a.s.} \Omega_e(s).$$

Since this holds with probability 1 over s, we have that

$$\Omega_e^N(s, n) \xrightarrow{n-a.s.} \Omega_e(s).$$

It follows that we can write

$$R_{e,i}^N(s, n) = e_i \circ k_i(s + n_i) - \Omega_e^N(s, n)$$
$$\xrightarrow{a.s.} R_{e,i}(s, n_i) := e_i \circ k_i(s + n_i) - \Omega_e(s)$$

H PROOF OF THEOREM 8

We will begin by showing that if $K \geq \mathrm{Var}(s) + C$ then $\{k_i^{-1}\} \in \mathcal{G}_K$.

For $e_i = k_i^{-1}$, we have that

$$\Omega_e^N(s, n) = \frac{1}{N} \sum_{i=1}^{N} s + n_i \xrightarrow{a.s.} s = \Omega_e^N(s)$$

$$R_i^N = s + n_i - \Omega_e(s, n) \xrightarrow{a.s.} n_i = R_{e,i}(n_i)$$

where the convergences follow from application of Kolmogorov's strong law, using the fact that $\mathrm{Var}(n_i) \leq C$ for all i. Satisfaction of condition 17 follows from the fact that $\mathrm{Var}_{s,n_i}(s + n_i) \leq C + \mathrm{Var}(s) \leq K$. Since s is a well-defined random variable, $\Omega_e(s) < \infty$ with probability 1, satisfying condition 18. It follows from the mutual independence of n_i and n_j that $R_{e,i}$ and $R_{e,j}$ satisfy condition 19. Condition 21 follows from the fact that $\mathbb{E}[n_i] = 0$ Condition 22 follows from $R_{e,i}$ being constant as a function of s.

It therefore follows that $\{k_i^{-1}\} \in \mathcal{G}_K$ for K sufficiently large.

We will next show that if $\{e_i\} \in \mathcal{G}_K$ then there exist a matrix α and vector β such that $e_i = \alpha k_i^{-1} + \beta$ for all i. Since e_i acts coordinate-wise, it moreover follows that α is diagonal.

First, we will show that each $e_i \circ k_i$ is affine, i.e. there exist potentially different α_i, β_i such that $e_i = \alpha_i k_i^{-1} + \beta_i$ for each i.

Then we will show that we must have $\alpha_i = \alpha_j$ and $\beta_i = \beta_j$ for all i, j.

To see that e_i is affine, we make use of that fact that $R_{e,i}$ is constant as a function of s. It follows that for any x and y

$$
\begin{aligned}
e_i \circ k_i(x + y) &= R_{e,i}(x) + \Omega_e(y) \\
&= R_{e,i}(x) + \Omega_e(0) + R_{e,i}(0) + \Omega_e(y) \\
&\quad - (R_{e,i}(0) + \Omega_e(0)) \\
&= e_i \circ k_i(x) + e_i \circ k_i(y) - e_i \circ k_i(0)
\end{aligned}
$$

It therefore follows that $e_i \circ k_i$ is affine, since if we define

$$
\begin{aligned}
L(x + y) &= e_i \circ k_i(x + y) - e_i \circ k_i(0) \\
&= (e_i \circ k_i(x) - e_i \circ k_i(0)) \\
&\quad + (e_i \circ k_i(y) - e_i \circ k_i(0)) \\
&= L(x) + L(y)
\end{aligned}
$$

then L is linear and we can write $e_i \circ k_i(x)$ as the sum of a linear function and a constant:

$$e_i \circ k_i(x) = L(x) + e_i \circ k_i(0)$$

Thus $e_i \circ k_i$ is affine, and we have some (diagonal) matrix α_i and vector β_i such that for any x

$$
\begin{aligned}
e_i \circ k_i(x) &= \alpha_i x + \beta_i \\
\implies e_i(x) &= \alpha_i k_i^{-1} x + \beta_i.
\end{aligned}
$$

Next we show that for the set of $\{e_i = \alpha_i k_i^{-1} + \beta_i\}$, it must be the case that each $\alpha_i = \alpha_j$ and $\beta_i = \beta_j$.

Observe that

$$
\begin{aligned}
\Omega_e^N(s, n) &= \frac{1}{N} \sum_{i=1}^{N} \alpha_i s + \alpha_i n_i + \beta_i \\
&= \left(\frac{1}{N} \sum_{i=1}^{N} \alpha_i \right) s + \frac{1}{N} \sum_{i=1}^{N} \beta_i + \frac{1}{N} \sum_{i=1}^{N} \alpha_i n_i
\end{aligned}
$$

$$\mathbb{E}_n[\Omega_e^N(s, n)] = \left(\frac{1}{N} \sum_{i=1}^{N} \alpha_i \right) s + \frac{1}{N} \sum_{i=1}^{N} \beta_i$$

Define

$$\alpha = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \alpha_i$$

$$\beta = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \beta_i$$

which exist by the assumption that $\Omega_e^N(s, n)$ converges as $N \to \infty$. Thus

$$
\begin{aligned}
\Omega_e(s) &= \alpha s + \beta \\
R_{e,i}(s, n_i) &= (\alpha_i - \alpha)s + \alpha_i n_i + \beta_i - \beta
\end{aligned}
$$

Now, suppose that there exist i and j such that such that $\alpha_i \neq \alpha_j$. It follows that

$$
\begin{aligned}
R_{e,i}(s, n_i) &= (\alpha_i - \alpha)s + \alpha_i n_i + \beta_i - \beta \\
R_{e,j}(s, n_j) &= (\alpha_j - \alpha)s + \alpha_j n_j + \beta_j - \beta
\end{aligned}
$$

There are two cases. If $\alpha_i \neq \alpha$, then $R_{e,i}(s, n_i)$ is not a constant function of s. But if $\alpha_i = \alpha$, then $\alpha_j \neq \alpha$ and so $R_{e,j}(s, n_j)$ is not a constant function of s. This is a contradiction, and so $\alpha_i = \alpha_j$ for all i, j.

Suppose similarly that there exist $\beta_i \neq \beta_j$. If $\beta_i \neq \beta$, then $\mathbb{E}[R_{e,i}(n_i)] = \beta_i - \beta$ which is non-zero. If $\beta_i = \beta$, then $\beta_j \neq \beta$ and so $\mathbb{E}[R_{e,j}(n_j)] = \beta_j - \beta$ is non-zero. This is a contradiction, and so $\beta_i = \beta_j$ for all i, j.

We have thus proven that set $\{e_i\} \in \mathcal{G}_K$ is of the form $e_i = \alpha k_i^{-1} + \beta$ for all i.