Supplementary Material for

Variational Autoencoders and Nonlinear ICA: A Unifying Framework

published at AISTATS 2020

A LEMMAS FOR THE EXPONENTIAL FAMILIES

We only consider univariate distributions in this section. The domain of the distributions is assumed to be $\mathbb{R}$, but all results hold if we replace $\mathbb{R}$ by an open set $\mathcal{Z} \subset \mathbb{R}$ whose Lebesgue measure is greater than 0.

A.1 Exponential family distributions

Definition 3 (Exponential family) *A univariate exponential family is a set of distributions whose probability density function can be written as*

$$p(x) = Q(x)Z(\boldsymbol{\theta})e^{\langle \mathbf{T}(x), \boldsymbol{\theta} \rangle} \tag{15}$$

where $\mathbf{T} : \mathbb{R} \to \mathbb{R}^k$ *is called the sufficient statistic,* $\boldsymbol{\theta} \in \mathbb{R}^k$ *is the natural parameter,* $Q : \mathbb{R} \to \mathbb{R}$ *the base measure and* $Z(\boldsymbol{\theta})$ *the normalization constant. The dimension* $k \in \mathbb{N} \setminus \{0\}$ *of the parameter is always considered to be* minimal*, meaning that we can't rewrite the density* p *to have the form (15) with a smaller* $k' < k$*. We call* k *the size of* p*.*

Lemma 1 *Consider an exponential family distribution with* $k \geq 2$ *components. If there exists* $\alpha \in \mathbb{R}^k$ *such that* $T_k(x) = \sum_{i=1}^{k-1} \alpha_i T_i(x) + \alpha_k$*, then* $\alpha = 0$*. In particular, the components of the sufficient statistic* $\mathbf{T}$ *are linearly independent.*

Proof: Suppose the components $(T_1, \dots, T_k)$ are not linearly independent. Then $\exists \boldsymbol{\alpha} \in \mathbb{R}^k \setminus \{0\}$ such that $\forall x \in \mathbb{R}$, $\sum_{i=1}^{k} \alpha_i T_i(x) = 0$. Suppose $\alpha_k \neq 0$ (up to rearrangement of the indices), then we can write T_k as a function of the remaining $T_i, i < k$, contradicting the minimality of k. □

A.2 Strongly exponential distributions

Definition 4 (Strongly exponential distributions) *We say that an exponential family distribution is* strongly exponential *if for any subset* $\mathcal{X}$ *of* $\mathbb{R}$ *the following is true:*

$$\left(\exists\, \boldsymbol{\theta} \in \mathbb{R}^k \,|\, \forall x \in \mathcal{X}, \langle \mathbf{T}(x), \boldsymbol{\theta} \rangle = const\right) \implies (l(\mathcal{X}) = 0 \text{ or } \boldsymbol{\theta} = 0) \tag{16}$$

where l *is the Lebesgue measure.*

In other words, the density of a strongly exponential distribution has almost surely the exponential component in its expression and can only be reduced to the base measure on a set of measure zero.

Example 1 *The strongly exponential condition is very general, and is satisfied by all the usual exponential family distributions like the Gaussian, Laplace, Pareto, Chi-squared, Gamma, Beta, etc.*

We will now give useful Lemmas that will be used in the proofs of the technical Theorems.

Lemma 2 *Consider a* strongly exponential *family distribution such that its sufficient statistic* $\mathbf{T}$ *is* differentiable *almost surely. Then* $T_i' \neq 0$ *almost everywhere on* $\mathbb{R}$ *for all* $1 \leq i \leq k$*.*

Proof: Suppose that p is strongly exponential, and let $\mathcal{X} = \cup_i \{x \in \mathbb{R}, T_i'(x) \neq 0\}$. Chose any $\boldsymbol{\theta} \in \mathbb{R}^k \setminus \{0\}$. Then $\forall x \in \mathcal{X}, \langle \mathbf{T}'(x), \boldsymbol{\theta} \rangle = 0$. By integrating, we find that $\langle \mathbf{T}(x), \boldsymbol{\theta} \rangle = \text{const}$. By hypothesis, this means that $l(\mathcal{X}) = 0$. □

Lemma 3 *Consider a strongly exponential distribution of size* $k \geq 2$ *with sufficient statistic* $\mathbf{T}(x) = (T_1(x), \dots, T_k(x))$*. Further assume that* $\mathbf{T}$ *is* differentiable *almost everywhere. Then there exist* k *distinct values* x_1 *to* x_k *such that* $(\mathbf{T}'(x_1), \dots, \mathbf{T}'(x_k))$ *are linearly independent in* $\mathbb{R}^k$*.*

Proof: Suppose that for any choice of such k points, the family $(\mathbf{T}'(x_1), \dots, \mathbf{T}'(x_k))$ is never linearly independent. That means that $\mathbf{T}'(\mathbb{R})$ is included in a subspace of $\mathbb{R}^k$ of dimension at most $k-1$. Let $\boldsymbol{\theta}$ a non zero vector that is orthogonal to $\mathbf{T}'(\mathbb{R})$. Then for all $x \in \mathbb{R}$, we have $\langle \mathbf{T}'(x), \boldsymbol{\theta} \rangle = 0$. By integrating we find that $\langle \mathbf{T}(x), \boldsymbol{\theta} \rangle = \text{const}$. Since this is true for all $x \in \mathbb{R}$ and for a $\boldsymbol{\theta} \neq 0$, we conclude that the distribution is not strongly exponential, which contradicts our hypothesis. □

Lemma 4 *Consider a strongly exponential distribution of size $k \geq 2$ with sufficient statistic $\mathbf{T}$. Further assume that $\mathbf{T}$ is* twice differentiable *almost everywhere. Then*

$$\dim\left(\text{span}\left(\left(T_i'(x), T_i''(x)\right)^T, 1 \leq i \leq k\right)\right) \geq 2 \tag{17}$$

almost everywhere on $\mathbb{R}$.

Proof: Suppose there exists a set $\mathcal{X}$ of measure greater than zero where (17) doesn't hold. This means that the vectors $[T_i'(x), T_i''(x)]^T$ are collinear for any i and for all $x \in \mathcal{X}$. In particular, it means that there exists $\alpha \in \mathbb{R}^k \setminus \{0\}$ s.t. $\sum_i \alpha_i T_i'(x) = 0$. By integrating, we get $\langle \mathbf{T}(x), \boldsymbol{\alpha} \rangle = \text{const}$, $\forall x \in \mathcal{X}$. Since $l(\mathcal{X}) > 0$, this contradicts equation (16). □

Lemma 5 *Consider n strongly exponential distributions of size $k \geq 2$ with respective sufficient statistics $\mathbf{T}_j = (T_{j,1}, \dots T_{j,k})$, $1 \leq j \leq n$. Further assume that the sufficient statistics are* twice differentiable. *Define the vectors $\mathbf{e}^{(j,i)} \in \mathbb{R}^{2n}$, such that $\mathbf{e}^{(j,i)} = \big(0, \dots, 0, T_{j,i}', T_{j,i}'', 0, \dots, 0\big)$, where the non-zero entries are at indices $(2j, 2j+1)$. Let $\mathbf{x} := (x_1, \dots, x_n) \in \mathbb{R}^n$ Then the matrix $\overline{e}(\mathbf{x}) := (\mathbf{e}^{(1,1)}(x_1), \dots, \mathbf{e}^{(1,k)}(x_1), \dots \mathbf{e}^{(n,1)}(x_n), \dots, \mathbf{e}^{(n,k)}(x_n))$ of size $(2n \times nk)$ has rank $2n$ almost everywhere on $\mathbb{R}^n$.*

Proof: It is easy to see that the matrix $\overline{e}(\mathbf{x})$ has at least rank n, because by varying the index j in $\mathbf{e}^{(j,i)}$ we change the position of the non-zero entries. By changing the index i, we change the component within the same sufficient statistic. Now fix j and consider the submatrix $\left[\mathbf{e}^{(j,1)}(x_j), \dots, \mathbf{e}^{(j,k)}(x_j)\right]$. By using Lemma 4, we deduce that this submatrix has rank greater or equal to 2 because its columns span a subspace of dimensions greater or equal to 2 almost everywhere on $\mathbb{R}$. Thus, we conclude that the rank of $\overline{e}(\mathbf{x})$ is $2n$ almost everywhere on $\mathbb{R}^n$. □

We will give now an example of an exponential family distribution that is not strongly exponential.

Example 2 *Consider an exponential family distribution with density function*

$$p(x) = e^{-x^2} Z(\boldsymbol{\theta}) \exp\left(\theta_1 \min(0, x) - \theta_2 \max(0, x)\right) \tag{18}$$

This density sums to 1 and $Z(\boldsymbol{\theta})$ is well defined. Yet, $\mathbf{T}(x) = (\min(0, x), -\max(0, x))$ is differentiable almost everywhere, but $T_1'(\mathbb{R}_+) = 0$ and $T_2'(\mathbb{R}_-) = 0$. It follows that p is not strongly exponential.

B PROOFS

B.1 Proof of Definition 2

Proposition 2 *The binary relations $\sim_A$ and $\sim_P$ are equivalence relations on Θ.*

The following proof applies to both $\sim_A$ and $\sim_P$ which we will simply denote by $\sim$.

It is clear that $\sim$ is reflexive and symmetric. Let $((\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda}), (\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}}), (\overline{\mathbf{f}}, \overline{\mathbf{T}}, \overline{\boldsymbol{\lambda}})) \in \Theta^3$, s.t. $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda}) \sim (\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}})$ and $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda}) \sim (\overline{\mathbf{f}}, \overline{\mathbf{T}}, \overline{\boldsymbol{\lambda}})$. Then $\exists A_1, A_2$ and $\mathbf{c}_1, \mathbf{c}_2$ s.t.

$$\begin{aligned}
\mathbf{T}(\mathbf{f}^{-1}(\mathbf{x})) &= A_1 \tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1}(\mathbf{x})) + \mathbf{c}_1 \text{ and} \\
\overline{\mathbf{T}}(\overline{\mathbf{f}}^{-1}(\mathbf{x})) &= A_2 \mathbf{T}(\mathbf{f}^{-1}(\mathbf{x})) + \mathbf{c}_2 \\
&= A_2 A_1 \tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1}(\mathbf{x})) + A_2 \mathbf{c}_1 + \mathbf{c}_2 \\
&= A_3 \tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1}(\mathbf{x})) + \mathbf{c}_3
\end{aligned} \tag{19}$$

and thus $(\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}}) \sim (\overline{\mathbf{f}}, \overline{\mathbf{T}}, \overline{\boldsymbol{\lambda}})$. □

B.2 Proof of Theorem 1

B.2.1 Main steps of the proof

The proof of this Theorem is done in three steps.

In the first step, we use a simple convolutional trick made possible by assumption (i), to transform the equality of observed data distributions into equality of noiseless distributions. In other words, it simplifies the noisy case into a noiseless case. This step results in equation (29).

The second step consists of removing all terms that are either a function of observations $\mathbf{x}$ or auxiliary variables $\mathbf{u}$. This is done by introducing the points provided by assumption (iv), and using $\mathbf{u}_0$ as a "pivot". This is simply done in equations (29)-(32).

The last step of the proof is slightly technical. The the goal is to show that the linear transformation is invertible thus resulting in an equivalence relation. This is where we use assumption (iii).

B.2.2 Proof

Step I We introduce here the volume of a matrix denoted $\operatorname{vol} A$ as the product of the singular values of A. When A is full column rank, $\operatorname{vol} A = \sqrt{\det A^T A}$, and when A is invertible, $\operatorname{vol} A = |\det A|$. The matrix volume can be used in the change of variable formula as a replacement for the absolute determinant of the Jacobian (Ben-Israel, 1999). This is most useful when the Jacobian is a rectangular matrix ($n < d$). Suppose we have two sets of parameters $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$ and $(\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}})$ such that $p_{\mathbf{f},\mathbf{T},\boldsymbol{\lambda}}(\mathbf{x}|\mathbf{u}) = p_{\tilde{\mathbf{f}},\tilde{\mathbf{T}},\tilde{\boldsymbol{\lambda}}}(\mathbf{x}|\mathbf{u})$ for all pairs $(\mathbf{x}, \mathbf{u})$. Then:

$$\int_{\mathcal{Z}} p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u}) p_{\mathbf{f}}(\mathbf{x}|\mathbf{z}) \mathrm{d}\mathbf{z} = \int_{\mathcal{Z}} p_{\tilde{\mathbf{T}},\tilde{\boldsymbol{\lambda}}}(\mathbf{z}|\mathbf{u}) p_{\tilde{\mathbf{f}}}(\mathbf{x}|\mathbf{z}) \mathrm{d}\mathbf{z} \tag{20}$$

$$\Rightarrow \int_{\mathcal{Z}} p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u}) p_{\varepsilon}(\mathbf{x} - \mathbf{f}(\mathbf{z})) \mathrm{d}\mathbf{z} = \int_{\mathcal{Z}} p_{\tilde{\mathbf{T}},\tilde{\boldsymbol{\lambda}}}(\mathbf{z}|\mathbf{u}) p_{\varepsilon}(\mathbf{x} - \tilde{\mathbf{f}}(\mathbf{z})) \mathrm{d}\mathbf{z} \tag{21}$$

$$\Rightarrow \int_{\mathcal{X}} p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{f}^{-1}(\bar{\mathbf{x}})|\mathbf{u}) \operatorname{vol} J_{\mathbf{f}^{-1}}(\bar{\mathbf{x}}) p_{\varepsilon}(\mathbf{x} - \bar{\mathbf{x}}) \mathrm{d}\bar{\mathbf{x}} = \int_{\mathcal{X}} p_{\tilde{\mathbf{T}},\tilde{\boldsymbol{\lambda}}}(\tilde{\mathbf{f}}^{-1}(\bar{\mathbf{x}})|\mathbf{u}) \operatorname{vol} J_{\tilde{\mathbf{f}}^{-1}}(\bar{\mathbf{x}}) p_{\varepsilon}(\mathbf{x} - \bar{\mathbf{x}}) \mathrm{d}\bar{\mathbf{x}} \tag{22}$$

$$\Rightarrow \int_{\mathbb{R}^d} \tilde{p}_{\mathbf{T},\boldsymbol{\lambda},\mathbf{f},\mathbf{u}}(\bar{\mathbf{x}}) p_{\varepsilon}(\mathbf{x} - \bar{\mathbf{x}}) \mathrm{d}\bar{\mathbf{x}} = \int_{\mathbb{R}^d} \tilde{p}_{\tilde{\mathbf{T}},\tilde{\boldsymbol{\lambda}},\tilde{\mathbf{f}},\mathbf{u}}(\bar{\mathbf{x}}) p_{\varepsilon}(\mathbf{x} - \bar{\mathbf{x}}) \mathrm{d}\bar{\mathbf{x}} \tag{23}$$

$$\Rightarrow (\tilde{p}_{\mathbf{T},\boldsymbol{\lambda},\mathbf{f},\mathbf{u}} * p_{\varepsilon})(\mathbf{x}) = (\tilde{p}_{\tilde{\mathbf{T}},\tilde{\boldsymbol{\lambda}},\tilde{\mathbf{f}},\mathbf{u}} * p_{\varepsilon})(\mathbf{x}) \tag{24}$$

$$\Rightarrow F[\tilde{p}_{\mathbf{T},\boldsymbol{\lambda},\mathbf{f},\mathbf{u}}](\omega) \varphi_{\varepsilon}(\omega) = F[\tilde{p}_{\tilde{\mathbf{T}},\tilde{\boldsymbol{\lambda}},\tilde{\mathbf{f}},\mathbf{u}}](\omega) \varphi_{\varepsilon}(\omega) \tag{25}$$

$$\Rightarrow F[\tilde{p}_{\mathbf{T},\boldsymbol{\lambda},\mathbf{f},\mathbf{u}}](\omega) = F[\tilde{p}_{\tilde{\mathbf{T}},\tilde{\boldsymbol{\lambda}},\tilde{\mathbf{f}},\mathbf{u}}](\omega) \tag{26}$$

$$\Rightarrow \tilde{p}_{\mathbf{T},\boldsymbol{\lambda},\mathbf{f},\mathbf{u}}(\mathbf{x}) = \tilde{p}_{\tilde{\mathbf{T}},\tilde{\boldsymbol{\lambda}},\tilde{\mathbf{f}},\mathbf{u}}(\mathbf{x}) \tag{27}$$

where:

- in equation (22), J denotes the Jacobian, and we made the change of variable $\bar{\mathbf{x}} = \mathbf{f}(\mathbf{z})$ on the left hand side, and $\bar{\mathbf{x}} = \tilde{\mathbf{f}}(\mathbf{z})$ on the right hand side.
- in equation (23), we introduced
$$\tilde{p}_{\mathbf{T},\boldsymbol{\lambda},\mathbf{f},\mathbf{u}}(\mathbf{x}) = p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{f}^{-1}(\mathbf{x})|\mathbf{u}) \operatorname{vol} J_{\mathbf{f}^{-1}}(\mathbf{x}) \mathbb{1}_{\mathcal{X}}(\mathbf{x}) \tag{28}$$
on the left hand side, and similarly on the right hand side.
- in equation (24), we used $*$ for the convolution operator.
- in equation (25), we used $F[.]$ to designate the Fourier transform, and where $\varphi_{\varepsilon} = F[p_{\varepsilon}]$ (by definition of the characteristic function).
- in equation (26), we dropped $\varphi_{\varepsilon}(\omega)$ from both sides as it is non-zero almost everywhere (by assumption (i)).

Equation (27) is valid for all $(\mathbf{x}, \mathbf{u}) \in \mathcal{X} \times \mathcal{U}$. What is basically says is that for the distributions to be the same after adding the noise, the noise-free distributions have to be the same. Note that $\mathbf{x}$ here is a general variable and we are actually dealing with the noise-free probability densities.

Step II By taking the logarithm on both sides of equation (27) and replacing $p_{\mathbf{T},\boldsymbol{\lambda}}$ by its expression from (7), we get:

$$\log \operatorname{vol} J_{\mathbf{f}^{-1}}(\mathbf{x}) + \sum_{i=1}^{n}(\log Q_i(f_i^{-1}(\mathbf{x})) - \log Z_i(\mathbf{u}) + \sum_{j=1}^{k} T_{i,j}(f_i^{-1}(\mathbf{x}))\lambda_{i,j}(\mathbf{u})) = \log \operatorname{vol} J_{\tilde{\mathbf{f}}^{-1}}(\mathbf{x}) + \sum_{i=1}^{n}(\log \tilde{Q}_i(\tilde{f}_i^{-1}(\mathbf{x})) - \log \tilde{Z}_i(\mathbf{u}) + \sum_{j=1}^{k} \tilde{T}_{i,j}(\tilde{f}_i^{-1}(\mathbf{x}))\tilde{\lambda}_{i,j}(\mathbf{u})) \quad (29)$$

Let $\mathbf{u}_0, \dots, \mathbf{u}_{nk}$ be the points provided by assumption (iv) of the Theorem, and define $\overline{\boldsymbol{\lambda}}(\mathbf{u}) = \boldsymbol{\lambda}(\mathbf{u}) - \boldsymbol{\lambda}(\mathbf{u}_0)$. We plug each of those $\mathbf{u}_l$ in (29) to obtain $nk+1$ such equations. We subtract the first equation for $\mathbf{u}_0$ from the remaining nk equations to get for $l = 1, \dots, nk$:

$$\langle \mathbf{T}(\mathbf{f}^{-1}(\mathbf{x})), \overline{\boldsymbol{\lambda}}(\mathbf{u}_l) \rangle + \sum_i \log \frac{Z_i(\mathbf{u}_0)}{Z_i(\mathbf{u}_l)} = \langle \tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1}(\mathbf{x})), \overline{\tilde{\boldsymbol{\lambda}}}(\mathbf{u}_l) \rangle + \sum_i \log \frac{\tilde{Z}_i(\mathbf{u}_0)}{\tilde{Z}_i(\mathbf{u}_l)} \quad (30)$$

Let L bet the matrix defined in assumption (iv), and $\tilde{L}$ similarly defined for $\tilde{\boldsymbol{\lambda}}$ ($\tilde{L}$ is not necessarily invertible). Define $b_l = \sum_i \log \frac{\tilde{Z}_i(\mathbf{u}_0) Z_i(\mathbf{u}_l)}{Z_i(\mathbf{u}_0) \tilde{Z}_i(\mathbf{u}_l)}$ and $\mathbf{b}$ the vector of all b_l for $l = 1, \dots, nk$. Expressing (30) for all points $\mathbf{u}_l$ in matrix form, we get:

$$L^T \mathbf{T}(\mathbf{f}^{-1}(\mathbf{x})) = \tilde{L}^T \tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1}(\mathbf{x})) + \mathbf{b} \quad (31)$$

We multiply both sides of (31) by the transpose of the inverse of L^T from the left to find:

$$\mathbf{T}(\mathbf{f}^{-1}(\mathbf{x})) = A \tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1}(\mathbf{x})) + \mathbf{c} \quad (32)$$

where $A = L^{-T}\tilde{L}$ and $\mathbf{c} = L^{-T}\mathbf{b}$.

Step III Now by definition of $\mathbf{T}$ and according to assumption (iii), its Jacobian exists and is an $nk \times n$ matrix of rank n. This implies that the Jacobian of $\tilde{\mathbf{T}} \circ \tilde{\mathbf{f}}^{-1}$ exists and is of rank n and so is A. We distinguish two cases:

- If $k = 1$, then this means that A is invertible (because A is $n \times n$).
- If $k > 1$, define $\bar{\mathbf{x}} = \mathbf{f}^{-1}(\mathbf{x})$ and $\mathbf{T}_i(\bar{x}_i) = (T_{i,1}(\bar{x}_i), \dots T_{i,k}(\bar{x}_i))$. According to Lemma 3, for each $i \in [1, \dots, n]$ there exist k points $\bar{x}_i^1, \dots, \bar{x}_i^k$ such that $(\mathbf{T}_i'(\bar{x}_i^1), \dots, \mathbf{T}_i'(\bar{x}_i^k))$ are linearly independent. Collect those points into k vectors $(\bar{\mathbf{x}}^1, \dots, \bar{\mathbf{x}}^k)$, and concatenate the k Jacobians $J_{\mathbf{T}}(\bar{\mathbf{x}}^l)$ evaluated at each of those vectors horizontally into the matrix $Q = (J_{\mathbf{T}}(\bar{\mathbf{x}}^1), \dots, J_{\mathbf{T}}(\bar{\mathbf{x}}^k))$ (and similarly define $\tilde{Q}$ as the concatenation of the Jacobians of $\tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1} \circ \mathbf{f}(\bar{\mathbf{x}}))$ evaluated at those points). Then the matrix Q is invertible (through a combination of Lemma 3 and the fact that each component of $\tilde{T}$ is univariate). By differentiating (32) for each $\mathbf{x}^l$, we get (in matrix form):
$$Q = A\tilde{Q} \quad (33)$$
The invertibility of Q implies the invertibility of A and $\tilde{Q}$.

Hence, (32) and the invertibility of A mean that $(\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}}) \sim (\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$.

Moreover, we have the following observations:

- the invertibility of A and L imply that $\tilde{L}$ is invertible,
- because the Jacobian of $\tilde{\mathbf{T}} \circ \tilde{\mathbf{f}}^{-1}$ is full rank and $\tilde{\mathbf{f}}$ is injective (hence its Jacobian is full rank too), $J_{\tilde{\mathbf{T}}}$ has to be full rank too, and $\tilde{T}_{i,j}'(z) \neq 0$ almost everywhere.
- the real equivalence class of identifiability may actually be narrower that what is defined by $\sim$, as the matrix A and the vector $\mathbf{c}$ here have very specific forms, and are functions of $\boldsymbol{\lambda}$ and $\tilde{\boldsymbol{\lambda}}$. □

B.2.3 Understanding assumption (iv) in Theorem 1

Let $\mathbf{u}^0$ be an arbitrary point in its support $\mathcal{U}$, and $h(\mathbf{u}) = \big(\lambda_{1,1}(\mathbf{u}) - \lambda_{1,1}(\mathbf{u}^0), \dots, \lambda_{n,k}(\mathbf{u}) - \lambda_{n,k}(\mathbf{u}^0)\big) \in \mathbb{R}^{nk}$. Saying that there exists nk distinct points $\mathbf{u}^1$ to $\mathbf{u}^{nk}$ (all different from $\mathbf{u}^0$) such that L is invertible is equivalent to saying that the vectors $\mathbf{h} := (h(\mathbf{u}^1), \dots, h(\mathbf{u}^{nk}))$ are linearly independent in $\mathbb{R}^{nk}$. Let's suppose for a second that for any such choice of points, these vectors are not linearly independent. This means that $h(\mathcal{U})$ is necessarily included in a subspace of $\mathbb{R}^{nk}$ of dimension at most $nk-1$. Such a subspace has measure zero in $\mathbb{R}^{nk}$. Thus, if $h(\mathcal{U})$ isn't included in a subset of measure zero in $\mathbb{R}^{nk}$, this can't be true, and there exists a set of points $\mathbf{u}^1$ to $\mathbf{u}^{nk}$ (all different from $\mathbf{u}^0$) such that L is invertible. This implies that as long as the $\lambda_{i,j}(\mathbf{u})$ are generated randomly and independently, then almost surely, $h(\mathcal{U})$ won't be included in any such subset with measure zero, and the assumption holds.

We next gvive a simple example where this assumption always holds. Suppose $n = 2$ and $k = 1$, and that the auxiliary variable is a positive scalar. Consider sources $z_i \sim \mathcal{N}(0, \lambda_i(u))$ that are distributed according to Gaussian distributions with zero mean and variances modulated as follows:

$$\lambda_1(u) = u \tag{34}$$
$$\lambda_2(u) = u^2 \tag{35}$$

Because the functions $u \mapsto u$ and $u \mapsto u^2$ are linearly independent (as functions), then for any choice of "pivot" point u_0, for instance $u_0 = 1$, and any choice of distinct non-zero scalars u_1 and u_2, the columns of the matrix $L := (\boldsymbol{\lambda}(u_1) - 1, \boldsymbol{\lambda}(u_2) - 1)$ are linearly independent, and the matrix is invertible.

B.3 Proof of Theorem 2

B.3.1 Main steps of the proof

The proof of this Theorem is done in two main steps.

The first step is to show that $\tilde{\mathbf{f}}^{-1} \circ \mathbf{f}$ is a pointwise function. This is done by showing that the product of any two distinct partial derivatives of any component is always zero. Along with invertibility, this means that each component depends exactly on one variable. This is where we use the two additional assumptions required by the Theorem.

In the second step, we plug the result of the first step in the equation that resulted from Theorem 1 (see equation (41)). The fact that $\mathbf{T}$, $\tilde{\mathbf{T}}$ and $\tilde{\mathbf{f}}^{-1} \circ \mathbf{f}$ are all pointwise functions implies that A is necessarily a permutation matrix.

B.3.2 Proof

Step I In this Theorem we suppose that $k \geq 2$. The assumptions of Theorem 1 hold, and so we have

$$\mathbf{T}(\mathbf{f}^{-1}(\mathbf{x})) = A\tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1}(\mathbf{x})) + \mathbf{c} \tag{36}$$

for an invertible $A \in \mathbb{R}^{nk \times nk}$. We will index A by four indices (i, l, a, b), where $1 \leq i \leq n, 1 \leq l \leq k$ refer to the rows and $1 \leq a \leq n, 1 \leq b \leq k$ to the columns. Let $\mathbf{v}(\mathbf{z}) = \tilde{\mathbf{f}}^{-1} \circ \mathbf{f}(\mathbf{z}) : \mathcal{Z} \to \mathcal{Z}$. Note that $\mathbf{v}$ is bijective because $\mathbf{f}$ and $\tilde{\mathbf{f}}$ are injective. Our goal is to show that $v_i(\mathbf{z})$ is a function of only one z_{j_i}, for all i. We will denote by $v_i^s := \frac{\partial v_i}{\partial z_s}(\mathbf{z})$, and $v_i^{st} := \frac{\partial^2 v_i}{\partial z_s \partial z_t}(\mathbf{z})$. For each $1 \leq i \leq n$ and $1 \leq l \leq k$, we get by differentiating (36) with respect to z_s:

$$\delta_{is} T'_{i,l}(z_i) = \sum_{a,b} A_{i,l,a,b} \tilde{T}'_{a,b}(v_a(\mathbf{z})) v_a^s(\mathbf{z}) \tag{37}$$

and by differentiating (37) with respect to $z_t, t > s$:

$$0 = \sum_{a,b} A_{i,l,a,b} \left(\tilde{T}'_{a,b}(v_a(\mathbf{z})) v_a^{s,t}(\mathbf{z}) \; + \tilde{T}''_{a,b}(v_a(\mathbf{z})) v_a^s(\mathbf{z}) v_a^t(\mathbf{z}) \right) \tag{38}$$

This equation is valid for all pairs $(s,t), t > s$. Define $\mathbf{B}_a(\mathbf{z}) := \big(v_a^{1,2}(\mathbf{z}), \dots, v_a^{n-1,n}(\mathbf{z})\big) \in \mathbb{R}^{\frac{n(n-1)}{2}}$, $\mathbf{C}_a(\mathbf{z}) := \big(v_a^1(\mathbf{z}) v_a^2(\mathbf{z}), \dots, v_a^{n-1}(\mathbf{z}) v_a^n(\mathbf{z})\big) \in \mathbb{R}^{\frac{n(n-1)}{2}}$, $M(\mathbf{z}) := (\mathbf{B}_1(\mathbf{z}), \mathbf{C}_1(\mathbf{z}), \dots, \mathbf{B}_n(\mathbf{z}), \mathbf{C}_n(\mathbf{z}))$, $\mathbf{e}^{(a,b)} :=$

$(0, \dots, 0, T'_{a,b}, T''_{a,b}, 0, \dots, 0) \in \mathbb{R}^{2n}$, such that the non-zero entries are at indices $(2a, 2a+1)$ and $\overline{e}(\mathbf{z}) := (\mathbf{e}^{(1,1)}(z_1), \dots, \mathbf{e}^{(1,k)}(z_1), \dots, \mathbf{e}^{(n,1)}(z_n), \dots, \mathbf{e}^{(n,k)}(z_n)) \in \mathbb{R}^{2n \times nk}$. Finally, denote by $A_{i,l}$ the (i,l)-th row of A. Then by grouping equation (38) for all valid pairs (s,t) and pairs (i,l) and writing it in matrix form, we get:

$$M(\mathbf{z})\overline{e}(\mathbf{z})A = 0 \tag{39}$$

Now by Lemma 5, we know that $\overline{e}(\mathbf{z})$ has rank $2n$ almost surely on $\mathcal{Z}$. Since A is invertible, it is full rank, and thus $\text{rank}(\overline{e}\,(\mathbf{z})A) = 2n$ almost surely on $\mathcal{Z}$. It suffices then to multiply by its pseudo-inverse from the right to get

$$M(\mathbf{z}) = 0 \tag{40}$$

In particular, $C_a(\mathbf{z}) = 0$ for all $1 \leq a \leq n$. This means that the Jacobian of $\mathbf{v}$ at each $\mathbf{z}$ has at most one non-zero entry in each row. By invertibility and continuity of $J_{\mathbf{v}}$, we deduce that the location of the non-zero entries are fixed and do not change as a function of $\mathbf{z}$. This proves that $\tilde{\mathbf{f}}^{-1} \circ \mathbf{f}$ is point-wise nonlinearity.

Step II Let $\overline{T}(\mathbf{z}) = \tilde{T}(\mathbf{v}(\mathbf{z})) + A^{-1}\mathbf{c}$. $\overline{\mathbf{T}}$ is a composition of a permutation and pointwise nonlinearity. Without any loss of generality, we assume that the permutation in $\overline{T}$ is the identity. Plugging this back into equation (36) yields:

$$\mathbf{T}(\mathbf{z}) = A\overline{\mathbf{T}}(\mathbf{z}) \tag{41}$$

Let $D = A^{-1}$. The last equation is valid for every component:

$$\overline{T}_{i,l}(z_i) = \sum_{a,b} D_{i,l,a,b} T_{a,b}(z_a) \tag{42}$$

By differentiating both sides with respect to z_s where $s \neq i$ we get

$$0 = \sum_{b} D_{i,l,s,b} T'_{s,b}(z_s) \tag{43}$$

By Lemma 1, we get $D_{i,l,s,b} = 0$ for all $1 \leq b \leq k$. Since (43) is valid for all l and all $s \neq i$, we deduce that the matrix D has a block diagonal form:

$$D = \begin{pmatrix} D_1 & & \\ & \ddots & \\ & & D_n \end{pmatrix} \tag{44}$$

We conclude that A has the same block diagonal form. Each block i transforms $\mathbf{T}_i(\mathbf{z})$ into $\overline{\mathbf{T}}_i(\mathbf{z})$, which achieves the proof. □

B.4 Proof of Theorem 3

B.4.1 Main steps of the proof

This proof uses concepts borrowed from differential geometry. A good reference is the monograph by Lee (2003).

By defining $\mathbf{v} = \mathbf{f}^{-1} \circ \tilde{\mathbf{f}}$, equation (32) implies that each function $T_i \circ v_i$ can be written as a separable sum, *i.e.* a sum of n maps where each map $h_{i,a}$ is function of only one component z_a.

Intuitively, since T_i is not monotonic, it admits a local extremum (supposed to be a minimum). By working locally around this minimum, we can suppose that it is global and attained at a unique point y_i. The smoothness condition on $\mathbf{v}$ imply that the manifold where $T_i \circ v_i$ is minimized has dimension $n-1$. This is where we need assumption (3.ii) of the Theorem.

On the other hand, because of the separability in the sum, each non constant $h_{i,k}$ (minimized as a consequence of minimizing $T_i \circ v_i$) introduces a constraint on this manifold that reduces its dimension by 1. That's why we can only have one non constant $h_{i,k}$ for each i.

B.4.2 Proof

In this Theorem we suppose that $k = 1$. For simplicity, we drop the exponential family component index: $T_i := T_{i,1}$. By introducing $\mathbf{v} = \mathbf{f}^{-1} \circ \tilde{\mathbf{f}}$ and $h_{i,a}(z_a) = A_{i,a}\tilde{T}_a(z_a) + \frac{c_i}{n}$ into equation (32), we can rewrite it as:

$$T_i(v_i(\mathbf{z})) = \sum_{a=1}^{n} h_{i,a}(z_a) \tag{45}$$

for all $1 \leq i \leq n$.

By assumption, $h_{i,a}$ is not monotonic, and so is T_i. So for each a, there exists $\tilde{y}_{i,a}$ where $h_{i,a}$ reaches an extremum, which we suppose is a minimum without loss of generality. This implies that $T_i \circ v_i$ reaches a minimum at $\tilde{\mathbf{y}}_i := (\tilde{y}_{i,1}, \dots, \tilde{y}_{i,n})$, which in turn implies that $y_i := v_i(\tilde{\mathbf{y}}_i)$ is a point where T_i reaches a local minimum. Let U be an open set centered around y_i, and let $\tilde{V} := v_i^{-1}[U]$ the preimage of U by v_i. Because v_i is continuous, $\tilde{V}$ is open in $\mathbb{R}^n$ and non-empty because $\tilde{\mathbf{y}}_i \in \tilde{V}$. We can then restrict ourselves to a cube $V \subset \tilde{V}$ that contains $\tilde{\mathbf{y}}_i$ which can be written as $V = V_1 \times \cdots \times V_n$ where each V_a is an open interval in $\mathbb{R}$.

We can chose U such that T_i has only one minimum that is reached at y_i. This is possible because $T_i' \neq 0$ almost everywhere by hypothesis. Similarly, we chose the cube V such that each $h_{i,a}$ either has only one minimum that is reached at $\tilde{y}_{i,a}$, or is constant (possible by setting $A_{i,a} = 0$). Define

$$m_i = \min_{\mathbf{z} \in V} T_i \circ v_i(\mathbf{z}) \in \mathbb{R} \tag{46}$$

$$\mu_{i,a} = \min_{z_a \in V_a} h_{i,a}(z_a) \in \mathbb{R} \tag{47}$$

for which we have $m_i = \sum_a \mu_{i,a}$.

Define the sets $C_i = \{\mathbf{z} \in V | T_i \circ v_i(\mathbf{z}) = m_i\}$, $\tilde{C}_{i,a} = \{\mathbf{z} \in V | h_{i,a}(z_a) = \mu_{i,a}\}$ and $\tilde{C}_i = \cap_a \tilde{C}_{i,a}$. We trivially have $\tilde{C}_i \subset C_i$. Next, we prove that $C_i \subset \tilde{C}_i$. Let $\mathbf{z} \in C_i$, and suppose $\mathbf{z} \notin \tilde{C}_i$. Then there exist an index k, $\varepsilon \in \mathbb{R}$ and $\tilde{\mathbf{z}} = (z_1, \dots, z_k + \varepsilon, \dots, z_n)$ such that $m_i = \sum_a h_{i,a}(z_a) > \sum_a h_{i,a}(\tilde{z}_a) \geq \sum_a \mu_{i,a} = m_i$ which is not possible. Thus $\mathbf{z} \in \tilde{C}_i$. Hence, $\tilde{C}_i = C_i$.

Since m_i is only reached at y_i, we have $C_i = \{\mathbf{z} \in V | v_i(\mathbf{z}) = y_i\}$. By hypothesis, v_i is of class $\mathcal{C}^1$, and its Jacobian is non-zero everywhere on V (by invertibility of $\mathbf{v}$). Then, by Corollary 5.14 in Lee (2003), we conclude that C_i is a smooth ($\mathcal{C}^1$) submanifold of co-dimension 1 in $\mathbb{R}^n$, and so is $\tilde{C}_i$ by equality.

On the other hand, if $h_{i,a}$ is not constant, then it reaches its minimum $\mu_{i,a}$ at only one point $\tilde{y}_{i,a}$ in V_a. In this case, $\tilde{C}_{i,a} = V_{[1,i-1]} \times \{\tilde{y}_{i,a}\} \times V_{[i+1,n]}$. Suppose that there exist two different indices $a \neq b$, such that $h_{i,a}$ and $h_{i,b}$ are not constant. Then $\tilde{C}_{i,a} \cap \tilde{C}_{i,b}$ is a submanifold of co-dimension 2. This would contradict the fact that the co-dimension of $\tilde{C}_i$ is 1.

Thus, exactly one of the $h_{i,a}$ is not constant for each i. This implies that the i-th row of matrix A has exactly one non-zero entry. The non-zero entry should occupy a different position in each row to guarantee invertibility, which proves that A is a scaled permutation matrix. Plugging this back into equation (32) implies that $\tilde{\mathbf{f}} \circ \mathbf{f}$ is a point-wise nonlinearity. □

B.5 Proof of Proposition 1

For simplicity, denote $Q(\mathbf{z}) := \prod_i Q_i(z_i)$ and $Z(\mathbf{u}) := \prod_i Z_i(\mathbf{u})$. Let A be an orthogonal matrix and $\tilde{\mathbf{z}} = A\mathbf{z}$ It is easy to check that $\tilde{\mathbf{z}} \sim p_{\tilde{\boldsymbol{\theta}}}(\tilde{\mathbf{z}}|\mathbf{u})$ where this new exponential family is defined by the quantities $\tilde{Q} = Q$, $\tilde{\mathbf{T}} = \mathbf{T}$, $\tilde{\boldsymbol{\lambda}} = A\boldsymbol{\lambda}$ and $\tilde{Z} = Z$. In particular, the base measure Q does not change when $Q_i(z_i) = 1$ or $Q_i(z_i) = e^{-z_i^2}$ because such a Q is a rotationally invariant function of $\mathbf{z}$. Further, we have

$$\langle \mathbf{z}, \boldsymbol{\lambda}(\mathbf{u}) \rangle = \langle A^T\tilde{\mathbf{z}}, \boldsymbol{\lambda}(\mathbf{u}) \rangle = \langle \tilde{\mathbf{z}}, A\boldsymbol{\lambda}(\mathbf{u}) \rangle = \langle \tilde{\mathbf{z}}, \tilde{\boldsymbol{\lambda}}(\mathbf{u}) \rangle \tag{48}$$

Finally let $\tilde{\mathbf{f}} = \mathbf{f} \circ A^T$, and $\tilde{\boldsymbol{\theta}} := \left(\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}}\right)$. We get:

$$p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}) \mathrm{d}\mathbf{z} \tag{49}$$

$$= \int p_{\varepsilon}(x - \mathbf{f}(\mathbf{z})) \frac{Q(\mathbf{z})}{Z(\mathbf{u})} \exp(\langle \mathbf{z}, \boldsymbol{\lambda}(\mathbf{u}) \rangle) \mathrm{d}\mathbf{z} \tag{50}$$

$$= \int p_{\varepsilon}(x - \tilde{\mathbf{f}}(\tilde{\mathbf{z}})) \frac{\tilde{Q}(\tilde{\mathbf{z}})}{\tilde{Z}(\mathbf{u})} \exp(\langle \tilde{\mathbf{z}}, \tilde{\boldsymbol{\lambda}}(\mathbf{u}) \rangle) \mathrm{d}\tilde{\mathbf{z}} \tag{51}$$

$$= p_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}|\mathbf{u}) \tag{52}$$

where in equation (51) we made the change of variable $\tilde{\mathbf{z}} = A\mathbf{z}$, and removed the Jacobian because it is equal to 1. We then see that it is not possible to distinguish between $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ based on the observed data distribution. □

B.6 Proof of Theorem 4

The loss (8) can be written as follows:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u}) - \mathrm{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}, \mathbf{u})\right) \tag{53}$$

If the family $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$ is large enough to include $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$, then by optimizing the loss over its parameter $\boldsymbol{\phi}$, we will minimize the KL term, eventually reaching zero, and the loss will be equal to the log-likelihood. The VAE in this case inherits all the properties of maximum likelihood estimation. In this particular case, since our identifiability is guaranteed up to equivalence classes, the consistency of MLE means that we converge to the equivalence class[7] (Theorem 1) of true parameter $\boldsymbol{\theta}^*$ *i.e.* in the limit of infinite data. □

C DISCRETE OBSERVATIONS

As explained in Maddison et al. (2016); Jang et al. (2016), categorical distributions can be viewed as a infinitesimal-temperature limit of continuous distributions. We can use this fact to extend our theory to discrete latent variables.

For example, let:

$$\mathbf{m} = \mathbf{f}(\mathbf{z}) \tag{54}$$

$$\mathbf{x} = \mathrm{sigmoid}((\mathbf{m} + \boldsymbol{\varepsilon})/T) \tag{55}$$

$$\forall \varepsilon_i \in \boldsymbol{\varepsilon} : \varepsilon_i \sim \mathrm{Logistic}(0, 1) \tag{56}$$

where sigmoid() is the element-wise sigmoid nonlinearity, and $T \in (0, \infty)$ is a temperature variable.

If we let T approach 0 from above, then:

$$\mathbf{x} \sim \mathrm{Bernoulli}(\mathbf{p}) \quad \text{with} \quad \mathbf{p} = \mathrm{sigmoid}(\mathbf{m}) \tag{57}$$

For proof that this holds, we refer to Maddison et al. (2016), appendix B.

The sigmoid$(\cdot/T)$ function is invertible, and the Logistic distribution has a probability density function that allows for deconvolution since its Fourier transform is non zero almost everywhere. As a result, for a given value of T, the distribution $p(\mathbf{x})$ has a one-to-one mapping to a distribution $p(\mathbf{m})$. This means that we can apply a small change to equations (20)-(27) and arrive at the same identifiability result. This example with a Bernoulli distribution can be extended to a categorical distribution with any number of components (Maddison et al., 2016; Jang et al., 2016).

D UNIDENTIFIABILITY OF GENERATIVE MODELS WITH UNCONDITIONAL PRIOR

In this section, we present two well-known proofs of unidentifiability of generative models. The first proof is simpler and considers factorial priors, which are widely-used in deep generative models and the VAE literature.

[7] this is easy to show: because true identifiability is one of the assumptions for MLE consistency, replacing it by identifiability up to equivalence class doesn't change the proof but only the conclusion.

The second proof is extremely general, and shows how any random vector can be transformed into independent components, in particular components which are standardized Gaussian. Thus, we see how in the general nonlinear case, there is little hope of finding the original latent variables based on the (unconditional, marginal) statistics of $\mathbf{x}$ alone.

D.1 Factorial priors

Let us start with factorial, Gaussian priors. In other words, let $\mathbf{z} \sim p_{\boldsymbol{\theta}}(\mathbf{z}) = N(\mathbf{0}, \mathbf{I})$. Now, a well-known result says that any orthogonal transformation of $\mathbf{z}$ has exactly the same distribution. Thus, we could transform the latent variable by any orthogonal transformation $\mathbf{z}' = M\mathbf{z}$, and cancel that transformation in $p(\mathbf{x}|\mathbf{z})$ (e.g. in the first layer of the neural network), and we would get exactly the same observed data (and thus obviously the same distribution of observed data) with $\mathbf{z}'$.

Formally we have

$$p_{\mathbf{z}'}(\boldsymbol{\xi}) = p_{\mathbf{z}}(M^T\boldsymbol{\xi})|\det M| = \frac{1}{(2\pi)^{d/2}} \exp(-\frac{1}{2}\|M^T\boldsymbol{\xi}\|^2) \tag{58}$$

$$= \frac{1}{(2\pi)^{d/2}} \exp(-\frac{1}{2}\|\boldsymbol{\xi}\|^2) = p_{\mathbf{z}}(\boldsymbol{\xi}) \tag{59}$$

where we have used the fact that the determinant of an orthogonal matrix is equal to unity.

This result applies easily to any factorial prior. For z_i of any distribution, we can transform it to a uniform distribution by $F_i(z_i)$ where F_i is the cumulative distribution function of z_i. Next, we can transform it into standardized Gaussian by $\Phi^{-1}(F_i(z_i))$ where Φ is the standardized Gaussian cdf. After this transformation, we can again take any orthogonal transformation without changing the distribution. And we can even transform back to the same marginal distributions by $F_i^{-1}(\Phi(.))$. Thus, the original latents are not identifiable.

D.2 General priors

The second proof comes from the theory of nonlinear ICA (Hyvärinen and Pajunen, 1999), from which the following Theorem is adapted.

Theorem 5 (Hyvärinen and Pajunen (1999)) *Let $\mathbf{z}$ be a d-dimensional random vector of any distribution. Then there exists a transformation $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d$ such that the components of $\mathbf{z}' := \mathbf{g}(\mathbf{z})$ are independent, and each component has a standardized Gaussian distribution. In particular, z'_1 equals a monotonic transformation of z_1.*

The proof is based on an iterative procedure reminiscent of Gram-Schmidt, where a new variable can always be transformed to be independent of any previously considered variables, which is why z_1 is essentially unchanged.

This Theorem means that there are infinitely many ways of defining independent components $\mathbf{z}$ that nonlinearly generated an observation $\mathbf{x}$. This is because we can first transform $\mathbf{z}$ any way we like and then apply the Theorem. The arbitrariness of the components is seen in the fact that we will always find that one arbitrary chosen variable in the transformation is one of the independent components. This is in some sense an alternative kind of indeterminacy to the one in the previous subsection.

In particular, we can even apply this Theorem on the observed data, taking $\mathbf{x}$ instead of $\mathbf{z}$. Then, in the case of factorial priors, just permuting the data variables, we would arrive at the conclusion that any of the x_i can be taken to be one of the independent components, which is absurd.

Now, to apply this theory in the case of a general prior on $\mathbf{z}$, it is enough to point out that we can transform any variable into independent Gaussian variables, apply any orthogonal transformation, then invert the transformation in the Theorem, and we get a nonlinear transformation $\mathbf{z}' = \mathbf{g}^{-1}(M\mathbf{g}(\mathbf{z}))$ which has exactly the same distribution as $\mathbf{z}$ but is a complex nonlinear transformation. Thus, no matter what the prior may be, by looking at the data alone, it is not possible to recover the true latents based an unconditional prior distribution, in the general nonlinear case.

E ALTERNATIVE FORMULATION OF THEOREM 1

Theorem 6 *Assume that we observe data sampled from a generative model defined according to (5)-(7), with parameters $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$. Assume the following holds:*

(i) The set $\{\mathbf{x} \in \mathcal{X} | \varphi_\varepsilon(\mathbf{x}) = 0\}$ has measure zero, where φ_ε is the characteristic function of the density p_ε defined in (6).

(ii) The mixing function $\mathbf{f}$ in (6) is injective.

(iii) The sufficient statistics $T_{i,j}$ in (7) are differentiable almost everywhere, and $T'_{i,j} \neq 0$ almost everywhere for all $1 \leq i \leq n$ and $1 \leq j \leq k$.

(iv) $\boldsymbol{\lambda}$ is differentiable, and there exists $\mathbf{u}_0 \in \mathcal{U}$ such that $J_{\boldsymbol{\lambda}}(\mathbf{u}_0)$ is invertible.

then the parameters $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$ are $\sim$-identifiable. Moreover, if there exists $(\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}})$ such that $p_{\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}}}(\mathbf{x}|\mathbf{u}) = p_{\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda}}(\mathbf{x}|\mathbf{u})$, then $\tilde{\mathbf{T}}$ and $\tilde{\boldsymbol{\lambda}}$ verify assumptions (iii) and (iv).

Proof: The start of the proof is similar to the proof of Theorem 1. When we get to equation (29):

$$\log \operatorname{vol} J_{\mathbf{f}^{-1}}(\mathbf{x}) + \sum_{i=1}^{n} (\log Q_i(f_i^{-1}(\mathbf{x})) - \log Z_i(\mathbf{u}) + \sum_{j=1}^{k} T_{i,j}(f_i^{-1}(\mathbf{x})) \lambda_{i,j}(\mathbf{u})) = \log \operatorname{vol} J_{\tilde{\mathbf{f}}^{-1}}(\mathbf{x}) + \sum_{i=1}^{n} (\log \tilde{Q}_i(\tilde{f}_i^{-1}(\mathbf{x})) - \log \tilde{Z}_i(\mathbf{u}) + \sum_{j=1}^{k} \tilde{T}_{i,j}(\tilde{f}_i^{-1}(\mathbf{x})) \tilde{\lambda}_{i,j}(\mathbf{u})) \quad (60)$$

we take the derivative of both sides with respect to $\mathbf{u}$ (assuming that $\tilde{\lambda}$ is also differentiable). All terms depending on $\mathbf{x}$ only disappear, and we are left with:

$$J_{\boldsymbol{\lambda}}(\mathbf{u})^T \mathbf{T}(\mathbf{f}^{-1}(\mathbf{x})) - \sum_i \nabla \log Z_i(\mathbf{u}) = J_{\tilde{\boldsymbol{\lambda}}}(\mathbf{u})^T \tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1}(\mathbf{x})) - \sum_i \nabla \log \tilde{Z}_i(\mathbf{u}) \quad (61)$$

By evaluating both sides at $\mathbf{u}_0$ provided by assumption (iv), and multiplying both sides by $J_{\boldsymbol{\lambda}}(\mathbf{u}_0)^{-T}$ (invertible by hypothesis), we find:

$$\mathbf{T}(\mathbf{f}^{-1}(\mathbf{x})) = A \tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1}(\mathbf{x})) + \mathbf{c} \quad (62)$$

where $A = J_{\boldsymbol{\lambda}}(\mathbf{u}_0)^{-T} J_{\tilde{\boldsymbol{\lambda}}}(\mathbf{u}_0)^T$ and $\mathbf{c} = \sum_i \nabla \log \frac{Z_i(\mathbf{u}_0)}{\tilde{Z}_i(\mathbf{u}_0)}$. The rest of the proof follows proof of Theorem 1, where in the last part we deduce that $J_{\tilde{\boldsymbol{\lambda}}}(\mathbf{u}_0)$ is invertible. □

F LINK BETWEEN MAXIMUM LIKELIHOOD AND TOTAL CORRELATION

Consider the noiseless case:

$$\mathbf{x} = \mathbf{f}(\mathbf{z}) \quad (63)$$

$$p(\mathbf{z}|\mathbf{u}) = \prod_i p_i(z_i|\mathbf{u}) \quad (64)$$

where the components of the latent variable are independent given the auxiliary variable $\mathbf{u}$. We can relate the log-likelihood of the data to the total correlation of the latent variables. To see this connection, let's use the change of variable formula in the expression of the log-likelihood:

$$\mathbb{E}_{p(\mathbf{x},\mathbf{u})} [\log p(\mathbf{x}|\mathbf{u})] = \mathbb{E}_{p(\mathbf{z},\mathbf{u})} \left[\sum_i \log p_i(z_i|\mathbf{u}) - \log |J_{\mathbf{f}}(\mathbf{z})| \right] \quad (65)$$

$$= -\mathbb{E}_{p(\mathbf{z},\mathbf{u})} [\log |J_{\mathbf{f}}(\mathbf{z})|] - \sum_i H(z_i|\mathbf{u}) \quad (66)$$

where $H(z_i|\mathbf{u})$ is the conditional differential entropy of z_i given $\mathbf{u}$. The same change of variable formula applied to $H(\mathbf{x}|\mathbf{u})$ yields:

$$H(\mathbf{x}|\mathbf{u}) = H(\mathbf{z}|\mathbf{u}) + \mathbb{E}_{p(\mathbf{z},\mathbf{u})}\left[\log|J_{\mathbf{f}}(\mathbf{z})|\right] \tag{67}$$

which we then use in the expression of the conditional total correlation:

$$\begin{aligned}\mathrm{TC}(\mathbf{z}|\mathbf{u}) &:= \sum_i H(z_i|\mathbf{u}) - H(\mathbf{z}|\mathbf{u}) \\ &= \sum_i H(z_i|\mathbf{u}) - H(\mathbf{x}|\mathbf{u}) + \mathbb{E}_{p(\mathbf{z},\mathbf{u})}\left[\log|J_{\mathbf{f}}(\mathbf{z})|\right]\end{aligned} \tag{68}$$

Putting equations (66) and (68) together, we get:

$$\mathbb{E}_{p(\mathbf{x},\mathbf{u})}[\log p(\mathbf{x}|\mathbf{u})] = -\mathrm{TC}(\mathbf{z}|\mathbf{u}) - H(\mathbf{x}|\mathbf{u}) \tag{69}$$

The last term in this equation is a function of the data only and is thus a constant. An algorithm which learns to maximize the data likelihood is decreasing the total correlation of the latent variable. The total correlation is measure of independence as it is equal to zero if and only if the components of the latent variable are independent. Thus, by using a VAE to maximize a lower bound on the data likelihood, we are trying to learn an estimate of the inverse of the mixing function that gives the most independent components.

G REMARKS ON PREVIOUS WORK

G.1 Previous work in nonlinear ICA

ICA by Time Contrastive Learning Time Contrastive Learning (TCL) introduced in Hyvärinen and Morioka (2016) is a method for nonlinear ICA based on the assumption that while the sources are independent, they are also non-stationary time series. This implies that they can be divided into known non-overlapping segments, such that their distributions vary across segments. The non-stationarity is supposed to be slow compared to the sampling rate, so that we can consider the distributions within each segment to be unchanged over time; resulting in a piecewise stationary distribution across segments. Formally, given a segment index $\tau \in \mathcal{T}$, where $\mathcal{T}$ is a finite set of indices, the distribution of the sources within that segment is modelled as an exponential family, which is in their notation:

$$\log p_\tau(\mathbf{s}) := \log p(\mathbf{s}|\text{segment} = \tau) = \sum_{j=1}^{d} \lambda_j(\tau) q(s_j) - \log Z(\tau) \tag{70}$$

where $q_{j,0}$ is a stationary baseline and q is the sufficient statistic for the exponential family of the sources (note that exponential families with different sufficient statistics for each source, or more than one sufficient statistic per source are allowed, but we focus on this simpler case here). Note that parameters λ_j depend on the segment index, indicating that the distribution of sources changes across segments. It follows from equation (11) that the observations are piece-wise stationary.

TCL recovers the inverse transformation $\mathbf{f}^{-1}$ by self-supervised learning, where the goal is to classify original data points against segment indices in a multinomial classification task. To this end, TCL employed a deep network consisting of a feature extractor $h(\mathbf{x}^{(i)}; \eta)$ with parameters η in the form of a neural network, followed by a final classifying layer (e.g. softmax). The theory of TCL, as stated in Theorem 1 of Hyvärinen and Morioka (2016), is premised on the fact that in order to optimally classify observations into their corresponding segments the feature extractor, $h(\mathbf{x}^{(i)}; \eta)$, must learn about the changes in the underlying distribution of latent sources. The theory shows that the method can learn the independent components up to transformations by sufficient statistics and a linear transformation, as in $\sim_A$ identifiability. It is further proposed that a linear ICA can recover the final A if the number of segments grows infinite and the segment distributions are random in a certain sense, but this latter assumption is unrealistic in applications where the number of segments is small. We also emphasize that our estimation method based on VAE is very different from such a self-supervised scheme.

ICA using auxiliary variables A more recent development in nonlinear ICA is given by Hyvärinen et al. (2019) where it is assumed that we observe data following a noiseless conditional nonlinear ICA case:

$$\mathbf{x} = \mathbf{f}(\mathbf{z}) \tag{71}$$

$$p(\mathbf{z}|\mathbf{u}) = \prod_i p_i(z_i|\mathbf{u}) \tag{72}$$

This formulation is so general that it subsumes previous models by Hyvärinen and Morioka (2016, 2017) in the sense of the data model. However, their estimation method is very different from TCL: They rely on a self-supervised binary discrimination task based on randomization to learn the unmixing function. More specifically, from a dataset of observations and auxiliary variables pairs $\mathcal{D} = \{\mathbf{x}^{(i)}, \mathbf{u}^{(i)}\}$, they construct a randomized dataset $\mathcal{D}^* = \{\mathbf{x}^{(i)}, \mathbf{u}^*\}$ where $\mathbf{u}^*$ is randomly drawn from the observed distribution of $\mathbf{u}$. To distinguish between both datasets, a deep logistic regression is used. The last hidden layer of the neural network is a feature extractor denoted $\mathbf{h}(\mathbf{x})$; like in TCL, the purpose of the feature extractor is therefore to extract the relevant features which will allow to distinguish between the two datasets. The identifiability results by Hyvärinen et al. (2019) have a lot of similarity to ours, and several of our proofs are inspired by them. However, we strengthen those results, while concentrating on the case of exponential family models. In particular, we show how any non-monotonic sufficient statistics for $k = 1$ leads to identifiability in Theorem 3, and also Theorem 2 generalizes the corresponding result (Theorem 2, case 2) in Hyvärinen et al. (2019). Again, their estimation method is completely different from ours.

G.2 Previous work on identifiability in VAEs

Our framework might look similar to semi-supervised learning methods in the VAE context, due to the inclusion of the auxiliary variable $\mathbf{u}$. However, the auxiliary variable $\mathbf{u}$ can play a more general role. For instance, in time-series, it can simply be the the time index or history; in audiovisual data, it can be either one of the modalities, where the other is used as an observation. More importantly, and to our knowledge, there is no proof of identifiability in the semi-supervised literature.

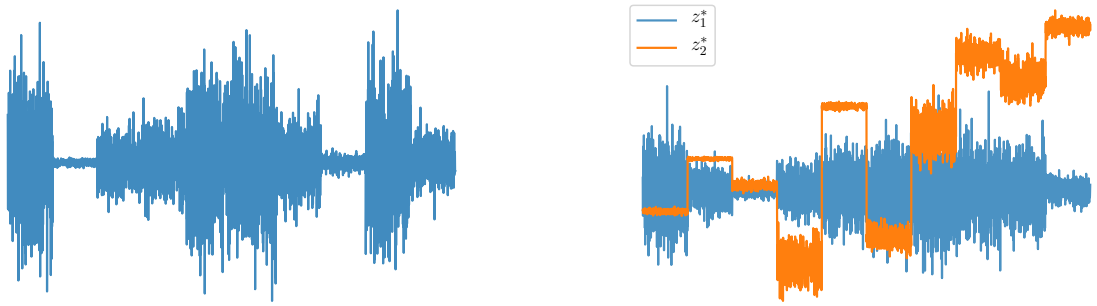
The question of identifiability, or lack of, in deep latent variable models especially VAEs has been tackled in work related to disentanglement. In Mathieu et al. (2018); Rolinek et al. (2018); Locatello et al. (2018) the authors show how isotropic priors lead to rotation invariance in the ELBO. We proved here (section 2.3 and supplementary material D) a much more general result: unconditional priors lead to unidentifiable models. These papers however focused on showcasing this problem, or how it can avoided in practice, and didn't provide alternative models that can be shown to be identifiable. This is what we try to achieve in this work, to provide a complementary analysis to previous research. Our proof of identifiability applies to the generative model itself, regardless of the estimation method. This is why we didn't focus in our analysis on the role of the encoder, which has been claimed to have a central role in some of the work cited above.

H SIMULATION DETAILS

H.1 Implementation detail for VAE experiments

We give here more detail on the data generation process for our simulations. The dataset is described in section 5.1. The conditioning variable $\mathbf{u}$ is the segment label, and its distribution is uniform on the integer set $[\![1, M]\!]$. Within each segment, the conditional prior distribution is chosen from the family (7), where $k = 1$, $T_{i,1}(z_i) = z_i^2$ and $Q_i(z_i) = 1$, and the true λ_i were randomly and independently generated across the segments and the components so that the variances have a uniform distribution on $[.5, 3]$. We sample latent variable $\mathbf{z}$ from these distribution, and then mix them using a 4-layer multi-layer perceptron (MLP). An example of what the sources look like is plotted in Figure 5a. We finally add small noise ($\sigma^2 = 0.01$) to the observations. When comparing to previous ICA methods, we omit this step, as these methods are for the noiseless case.

For the decoder (6), we chose $p_\varepsilon = \mathcal{N}\left(0, \sigma^2 I\right)$ a zero mean Gaussian, where the scalar σ^2 controls the noise level. We fix the noise level $\sigma^2 = 0.01$. As for the inference model, we let $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u}) = \mathcal{N}\left(\mathbf{z}|\mathbf{g}(\mathbf{x}, \mathbf{u}; \phi_{\mathbf{g}}), \mathbf{diag}\, \boldsymbol{\sigma}^2(\mathbf{x}, \mathbf{u}; \phi_{\boldsymbol{\sigma}})\right)$ be a multivariate Gaussian with a diagonal covariance. The functional parameters of the decoder ($\mathbf{f}$) and the inference model ($\mathbf{g}$, $\boldsymbol{\sigma}^2$) as well as the conditional prior ($\boldsymbol{\lambda}$) are chosen to be MLPs, where the dimension of the hidden layers is varied between 10 and 200, the activation function is a leaky ReLU, and the number of layers is



(a) (b)

Figure 5: Visualization of various sources following the generative distribution detailed in equation (7). (a) single source with segment modulated variance; (b) two sources where the mean of the second source, z_2^*, is significantly modulated as a function of the segment, thus potentially serving to greatly facilitate the surrogate classification task performed in TCL.

chosen from $\{3, 4, 5, 6\}$. Mini-batches are of size 64, and the learning rate of the Adam optimizer is chosen from $\{0.01, 0.001\}$. We also use a scheduler to decay the learning rate as a function of epochs.

To implement the VAE, we followed Kingma and Welling (2013). We made sure the range of the hyperparameters (mainly number of layers and dimension of hidden layers) of the VAE is large enough for it to be comparable in complexity to our method (which has the extra λ network to learn). To implement a β-VAE, we followed the instructions of Higgins et al. (2016) for the choice of hyperparameter β, which was chosen in the set $[1, 45]$. Similarly, we followed Chen et al. (2018) for the choice of the hyperparameters α, β and γ when implementing a β-TC-VAE: we chose $\alpha = \gamma = 1$ and β was chosen in the set $[1, 35]$.

H.2 Description of significant mean modulated data

Here, we generated non-stationary 2D data from a modified dataset as follows: $\mathbf{z}^*|u \sim \mathcal{N}(\boldsymbol{\mu}(u), \mathbf{diag}(\boldsymbol{\sigma}^2(u))$ where u is the segment index, $\mu_1(u) = 0$ for all u and $\mu_2(u) = \alpha\gamma(u)$ where $\alpha \in \mathbb{R}$ and γ is a permutation. Essentially, the mean of the second source, z_2^*, is significantly modulated by the segment index. An example is plotted in Figure 5b. The variance $\boldsymbol{\sigma}^2(u)$ is generated randomly and independently across the segments. We then mix the sources into observations $\mathbf{x}$ such that $x_1 = \mathbf{MLP}(z_1, z_2)$ and $x_2 = z_2^*$, thus preserving the significant modulation of the mean in x_2. We note that this is just one of many potential mappings from $\mathbf{z}$ to $\mathbf{x}$ which could have been employed to yield significant mean modulation in x_2 across segments. TCL learns to unmix observations, $\mathbf{x}$, by solving a surrogate classification task. Formally, TCL seeks to train a deep network to accurately classify each observation into its corresponding segment. As such, the aforementioned dataset is designed to highlight the following limitation of TCL: due to its reliance on optimizing a self-supervised objective, it can fail to recover latent variables when the associated task is too easy. In fact, by choosing a large enough value of the separation parameter α (in our experiments $\alpha = 2$), it is possible to classify samples by looking at the mean of x_2.

I FURTHER EXPERIMENTS

I.1 Additional general nonlinear ICA experiments

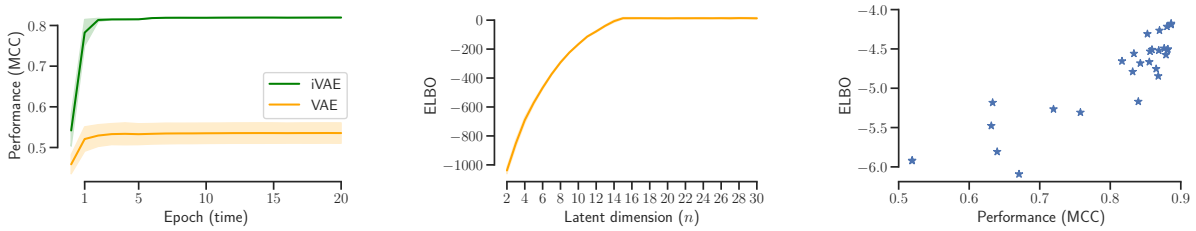
As discussed in section 3.4, our estimation method has many benefits over previously proposed self-supervised nonlinear ICA methods: it allows for dimensionality reduction, latent dimension selection based on the ELBO as a cross validation metric, and solving discrete ICA. We performed a series of simulations to test these claims.

Discrete observations To further test the capabilities of our method, we tested it on discrete data, and compared its identifiability performance to a vanilla VAE. The dimensions of the data and latents are $d = 100$ and

$n = 10$. The results are shown in Figure 6a and proves that our method is capable of performing discrete ICA.

Dimensionality selection and reduction The examples in section 5.1 already showcased dimensionality reduction. In Figure 2b for example, we have a mismatch between the dimensions of the latents and observations. In real world ICA applications, we usually don't know the dimension of the latents beforehand. One way to guess it is to use the ELBO as a proxy to select the dimension. Our method enables this when compared to previous nonlinear ICA methods like TCL (Hyvärinen and Morioka, 2016). This is showcased in Figure 6b, where the real dimensions of the simulated data are $d^* = 80$ and $n^* = 15$, and we run multiple experiments where we vary the latent dimensions between 2 and 40. We can see that the ELBO can be a good proxy for dimension selection, since it has a "knee" around the right value of dimension.

Hyperparameter selection One important benefit of the proposed method is that it seeks to optimize an objective function derived from the marginal log-likelihood of observations. As such, it follows that we may employ the ELBO to perform hyperparameter selection. To verify this claim, we run experiments for various distinct choices of hyperparameters (for example the dimension of hidden layers, number of hidden layers in the estimation network, learning rate, nonlinearities) on a synthetic dataset. Results are provided in Figure 6c which serves to empirically demonstrate that the ELBO is indeed a good proxy for how accurately we are able to recover the true latent variables. In contrast, alternative methods for nonlinear ICA, such as TCL, do not provide principled and reliable proxies which reflect the accuracy of estimated latent sources.



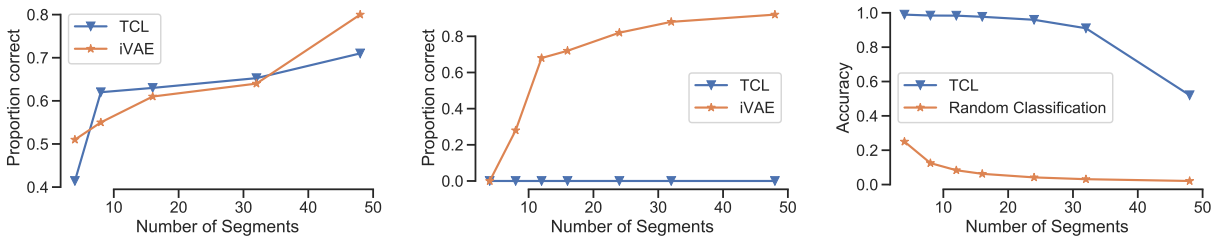
(a) iVAE for discrete ICA (b) Dimensionality selection (c) Hyperparameter selection

Figure 6: (a) Performance of iVAE and VAE on discrete ICA task. (b) Evolution of the post training ELBO as a function of the latent dimension. The real dimension of the data is $d^* = 80$ and the real dimension of the latent space is $n^* = 15$. We observe an elbow at around 15, thus successfully guessing the real dimension. (c) ELBO as a function of the performance. Each star is an experiment run for a different set of hyperparameters.

I.2 Additional causality experiments for comparison to TCL

Setup The data generation process used by (Monti et al., 2019, Section 4) is similar to the one we described in section 5.1, with the difference that the mixing should be in such a way that we get an acyclic causal relationship between the observations. This can be achieved by ensuring weight matrices in the mixing network are all lower-triangular, thereby introducing acyclic causal structure over observations.

Experiments on "normal" simulated data We seek to compare iVAE and TCL in the context of causal discovery, as described in Section 5.2. Such an approach involves a two-step procedure whereby first either TCL or iVAE are employed to recover latent disturbances, followed by a series of independence tests. Throughout all causal discovery experiments we employ HSIC as a general test of statistical independence (Gretton et al., 2005). When comparing iVAE and TCL in this setting we report the proportion of times the correct causal direction is reported. It is important to note that the aforementioned testing procedure can produce one of three decisions: $x_1 \to x_2$, $x_2 \to x_1$ or a third decision which states that no acyclic causal direction can be determined. The first two outcomes correspond to identifying causal structure and will occur when we fail to reject the null hypothesis in only one of the four tests. Whereas the third decision (no evidence of acyclic causal structure) will be reported when either there is evidence to reject the null in all four tests or we fail to reject the null more than once. Typically, this will occur if the nonlinear unmixing has failed to accurately recover the true latent sources. The results are reported in Figure 7a where we note that both TCL and iVAE perform comparably.



(a) (b) (c)

Figure 7: (*a*) Performance of nonlinear causal discovery for "normal" data, as described in Section I.2 when iVAE or TCL are employed to recover latent disturbances. (*b*) Similarly, but when underlying sources display significant mean modulation across segments, making them easy to classify. (*c*) Classification accuracy of TCL when applied on data displaying significant mean modulation. We note that the accuracy of TCL is significantly above a random classifier, indicating that the surrogate classification problem employed in TCL training has been effectively optimized.

Experiments on significant mean modulated data As a further experiment, we consider causal discovery in the scenario where one or both of the underlying sources demonstrate a significant mean modulation as shown in Figure 5. In such a setting the surrogate classification problem which is solved as part of TCL training becomes significantly easier, to the extent that TCL no longer needs to learn an accurate representation of the log-density of sources within each segment. This is to the detriment of TCL as it implies that it cannot accurately recover latent sources and therefore fails at the task of causal discovery. This can be seen in Figure 7b where TCL based causal discovery fails whereas iVAE continues to perform well. This is a result of the fact that iVAE directly optimizes the log-likelihood as opposed to a surrogate classification problem. Moreover, Figure 7c visualizes the mean classification accuracy for TCL as a function of the number of segments. We note that TCL consistently obtains classification accuracy that are significantly better that random classification. This provides evidence that the poor performance of TCL in the context of data with significant mean modulations is not a result of sub-optimal optimisation but are instead a a negative consequence of TCL's reliance on solving a surrogate classification problem to perform nonlinear unmixing.

I.3 Real data experiments

Hippocampal fMRI data Here we provide further details relating to the resting-state Hippocampal data provided by Poldrack et al. (2015) and studied in Section 5.2, closely following the earlier causal work using TCL by Monti et al. (2019). The data corresponds to daily fMRI scans from a single individual (Caucasian male, aged 45) collected over a period of 84 successive days. We consider data collected from each day as corresponding to a distinct segment, encoded in $\mathbf{u}$. Within each day 518 BOLD observations are provided across the following six brain regions: perirhinal cortex (PRc), parahippocampal cortex (PHc), entorhinal cortex (ERc), subiculum (Sub), CA1 and CA3/Dentate Gyrus (DG).

I.4 Additional visualisations for comparison to VAE variants

As a further visualization we show in Figures 8 and 9 the recovered latents for VAE and iVAE ; we sampled a random (contiguous) subset of the sources from the dataset, and compared them to the recovered latents (after inverting any permutation in the components). We can see that iVAE has an excellent estimation of the original sources compared to VAE (other models were almost indistinguishable from vanilla VAE).

J ACKNOWLEDGEMENT

I.K. and R.P.M. were supported by the Gatsby Charitable Foundation. A.H. was supported by a Fellowship from CIFAR, and from the DATAIA convergence institute as part of the "Programme d'Investissement d'Avenir", (ANR-17-CONV-0003) operated by Inria.

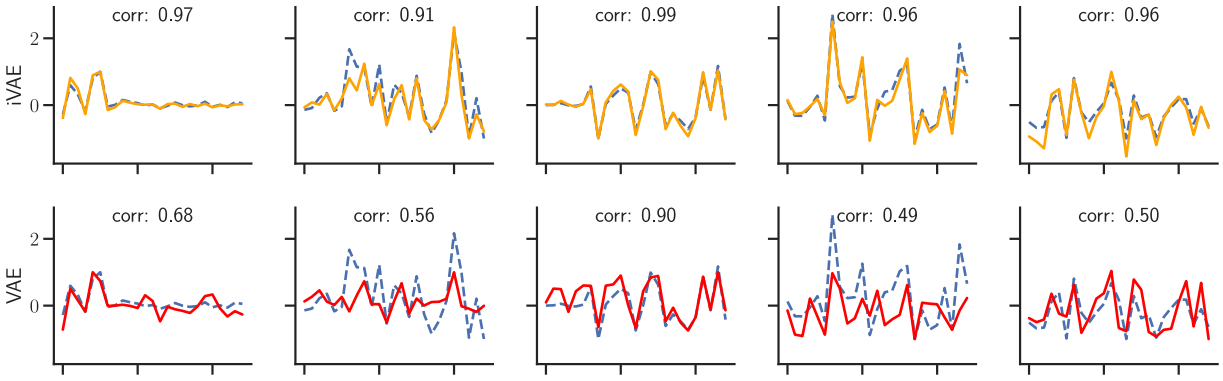


Figure 8: Comparison of the recovered latents of our model to the latents recovered by a vanilla VAE. The dashed blue line is the true source signal, and the recovered latents are in solid coloured lines. We also reported the correlation coefficients for every (source, latent) pair.

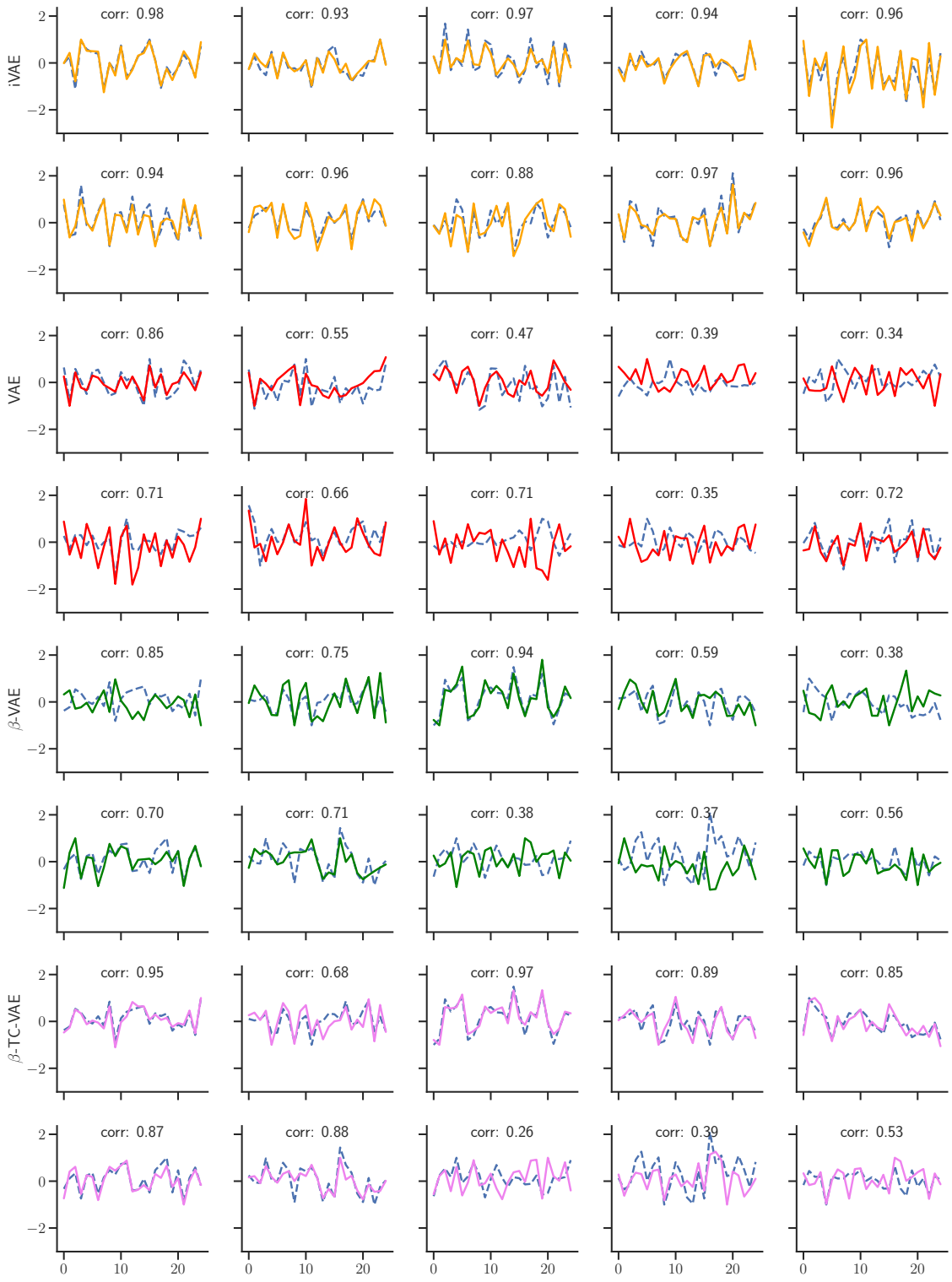


Figure 9: Comparison of the recovered latents of our model to the latents recovered by a vanilla VAE, a β-VAE and a β-TC-VAE, where the dimension of the data is $d = 40$, and the dimension of the latents is $n = 10$, the number of segments is $M = 40$ and the number of samples per segment is $L = 4000$. The dashed blue line is the true source signal, and the recovered latents are in solid coloured lines. We reported the correlation coefficients for every (source, latent) pair. We can see that iVAE have an excellent estimation of the original sources compared to the other models.