Learning Bijective Feature Maps for Linear ICA

Alexander Camuto*,1,3
Brooks Paige2,3
Chris Holmes1,3
Matthew Willetts*,1,3
Stephen Roberts1,3

1 University of Oxford
2 University College London
3 Alan Turing Institute

Abstract
Separating high-dimensional data like images into independent latent factors, i.e independent component analysis (ICA), remains an open research problem. As we show, existing probabilistic deep generative models (DGMs), which are tailor-made for image data, underperform on non-linear ICA tasks. To address this, we propose a DGM which combines bijective feature maps with a linear ICA model to learn interpretable latent structures for high-dimensional data. Given the complexities of jointly training such a hybrid model, we introduce novel theory that constrains linear ICA to lie close to the manifold of orthogonal rectangular matrices, the Stiefel manifold. By doing so we create models that converge quickly, are easy to train, and achieve better unsupervised latent factor discovery than flow-based models, linear ICA, and Variational Autoencoders on images.

1 Introduction
In linear Independent Component Analysis (ICA), data is modelled as having been created from a linear mixing of independent latent sources (Cardoso 1989a,b, 1997; Comon 1994). The canonical problem is blind source separation; the aim is to estimate the original sources of a mixed set of signals by learning an unmixing matrix, which when multiplied with data recovers the values of these sources. While linear ICA is a powerful approach to unmix signals like sound (Everson & Roberts 2001), it has not been as effectively developed for learning compact representations of high-dimensional data like images, where assuming linearity is limiting. Non-linear ICA methods, which assume non-linear mixing of latents, offer better performance on such data.

In particular, flow-based models have been proposed as a non-linear approach to square ICA, where we assume the dimensionality of our latent source space is the same as that of our data (Deco & Brauer 1995; Dinh et al. 2015). Flows parameterise a bijective mapping between data and a feature space of the same dimension and can be trained via maximum likelihood for a chosen base distribution in that space. While these are powerful generative models, for image data one typically wants fewer latent variables than the number of pixels in an image. In such situations, we wish to learn a non-square (dimensionality-reducing) ICA representation.

In this work, we highlight the fact that existing probabilistic deep generative models (DGMs), in particular Variational Autoencoders (VAEs), underperform on non-linear ICA tasks. As such there is a real need for a probabilistic DGM that can perform these tasks. To address this we propose a novel methodology for performing non-square non-linear ICA using a model, termed Bijecta, with two jointly trained parts: a highly-constrained non-square linear ICA model, operating on a feature space output by a bijective flow. The bijective flow is tasked with learning a representation for which linear ICA is a good model. We find that such a model fails to converge when trained naively with no constraints. To ensure convergence, we introduce novel theory for the parameterisation of decorrelating, non-square ICA matrices that lie close to the Stiefel manifold (Stiefel 1935), the space of orthonormal rectangular matrices. We use this result to introduce a novel non-square linear ICA model, operating on a feature space output by a bijective flow. The bijective flow is tasked with learning a representation for which linear ICA is a good model. It is as if we are learning the data for our ICA model.

We find that such a model fails to converge when trained naively with no constraints. To ensure convergence, we introduce novel theory for the parameterisation of decorrelating, non-square ICA matrices that lie close to the Stiefel manifold (Stiefel 1935), the space of orthonormal rectangular matrices. We use this result to introduce a novel non-square linear ICA model, operating on a feature space output by a bijective flow. The bijective flow is tasked with learning a representation for which linear ICA is a good model. It is as if we are learning the data for our ICA model.

*Equal Contribution.
Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).
2 Background

2.1 Independent Component Analysis

The goal of ICA is to learn a set of statistically independent sources that ‘explain’ our data. ICA is a highly diverse modelling paradigm with numerous variants: learning a mapping vs learning a model, linear vs non-linear, different loss functions, different generative models, and a wide array of methods of inference (Cardoso 1989a; Mackay 1996; Lee et al. 2000).

Here, we specify a generative model and find point-wise maximum likelihood estimates of model parameters in the manner of (Mackay 1996; Cardoso 1997). Concretely, we have a model with latent sources \( s \in \mathcal{S} = \mathbb{R}^{d_s} \) generating data \( x \in \mathcal{X} = \mathbb{R}^{d_x} \), with \( d_s \leq d_x \). The linear ICA generative model factorises as

\[
p(x, s) = p(x|s)p(s), \quad p(s) = \prod_{i=1}^{d_s} p(s_i),
\]

where \( p(s) \) is a set of independent distributions appropriate for ICA. In linear ICA, where all mappings are simple matrix multiplications, the sources cannot be Gaussian distributions. Recall that we are mixing our sources to generate our data: A linear mixing of Gaussian random variables is itself Gaussian, so unmixing is impossible (Lawrence & Bishop 2000). To be able to unmix, to break this symmetry, we can choose any heavy-tailed or light-tailed non-Gaussian distribution as our prior \( p(s) \) that gives us axis alignment and independence between sources. A common choice is the family of generalised Gaussian distributions,

\[
p(s_i) = GG(s_i|\mu, \alpha, \rho)
\]

\[
p(s_i) = \frac{\rho}{2\alpha\Gamma(1/\rho)} \exp \left( -\frac{|s_i - \mu|}{\alpha} \right)^\rho
\]

with mean \( \mu \), scale \( \alpha \) and shape \( \rho \). For \( \rho = 2 \) we recover the Normal distribution, and for \( \rho = 1 \) we have the (heavy-tailed) Laplace. As \( \rho \to \infty \) the distribution becomes increasingly sub-Gaussian, tending to a uniform distribution. As such, the generalised Gaussian is a flexible framework for specifying ICA-appropriate priors as it allows for the specification of a sub or super Gaussian distribution by way of a single parameter: \( \rho \).

2.2 Manifolds for the unmixing matrix \( A^+ \)

In linear ICA we want to find the linear mapping \( A^+ \) resulting in maximally independent sources. This is more onerous than merely finding decorrelated sources, as found by principal component analysis (PCA).

When learning a linear ICA model we typically have the mixing matrix \( A \) as the (pseudo)inverse of the unmixing matrix \( A^+ \) and focus on the properties of \( A^+ \) to improve convergence. \( A^+ \) linearly maps from the data-space \( \mathcal{X} \) to the source space \( \mathcal{S} \). It can be decomposed into two linear operations. First we whiten the data such that each component has unit variance and these components are mutually uncorrelated. We then apply an orthogonal transformation and a scaling operation (Hyvärinen et al. 2001, §6.34) to ‘rotate’ the whitened data into a set of coordinates where the sources are independent and decorrelated. Whitening on its own is not sufficient for ICA — two sources can be uncorrelated and dependent (see Appendix A).

Thus we can write the linear ICA unmixing matrix as

\[
A^+ = \Phi RW
\]

where \( W \in \mathbb{R}^{d_s \times d_s} \) is our whitening matrix, \( R \in \mathbb{R}^{d_x \times d_x} \) is an orthogonal matrix and \( \Phi \in \mathbb{R}^{d_s} \) is a diagonal matrix. Matrices that factorise this way are known as the decorrelating matrices (Everson & Roberts, 1999): members of this family decorrelate through \( W \), and \( \Phi R \) ensures that sources are statistically independent, not merely uncorrelated. The optimal ICA unmixing matrix is the decorrelating matrix that decorrelates and gives independence.

2.3 Flows

Flows are models that stack numerous invertible changes of variables. One specifies a simple base distribution and learns a sequence of (invertible) transforms to construct new distributions that assign high probability to observed data. Given a variable \( z \in \mathcal{Z} = \mathbb{R}^{d_z} \),

Figure 1: Here we take a dSprites heart and, using a randomly sampled affine transformation, move it around a black background (a). The underlying sources of the dataset are affine transformations of the heart. In (b-c) images in the center correspond to the origin of the learnt source space. Images on either side correspond to linearly increasing values along one of the learnt latent sources whilst the other source remains fixed. Bijecta (c) has learned affine transformations as sources (white diagonals), whereas a VAE (with ICA-appropriate prior) (b) has learned non linear transforms (white curves). The VAE has not discovered the underlying latent sources.
we specify the distribution over data $x$ as
\[ p(x) = p(z) \left| \det \frac{\partial f^{-1}}{\partial z} \right| , \quad (3) \]
where $f$ is a bijection from $Z \to X$, i.e $\mathbb{R}^{d_z} \to \mathbb{R}^{d_x}$, and
\[ p(z) \]
is the base distribution over the latent $z$ [Rezende & Mohamed, 2015; Papamakarios et al. 2019].

For more flexible distributions for $x$, we specify $x$ through a series of composed functions, from our simple initial $p$ into a more complex multi-modal distribution; for example for a series of $K+1$ mappings, $z = f_K \circ ... \circ f_0(x)$. By the properties of determinants under function composition
\[ p(x) = p(z_K) \prod_{i=0}^{K} \left| \det \frac{\partial f_{i-1}}{\partial z_{i+1}} \right| , \quad (4) \]
where $z_{i+1}$ denotes the variable resulting from the transformation $f_i(z_i)$, $p(z_K)$ defines a density on the $K^{\text{th}}$, and the bottom most variable is our data ($z_0 = x$).

Computing the determinant of the Jacobian ($\det \frac{\partial f}{\partial z}$) in Eq. (3) can be prohibitively costly, especially when composing multiple functions as in Eq. (1). To address this, flows use coupling layers that enforce a lower triangular Jacobian such that the determinant of the Jacobian is simply the product of its diagonal elements. We use recently proposed coupling layers based on rational quadratic splines (RQS) to enforce this lower triangular structure [Durkan et al., 2019]. They form highly flexible flows that typically require fewer composed mappings to achieve good performance relative to other coupling layers. See Appendix [I] for details.

3 Non-Square ICA using Flows

Variational Autoencoders seem like a natural fit for learning a compressed set of statistically independent latent variables [Kingma & Welling, 2014; Rezende et al., 2014]. It seems natural to train a VAE with an appropriate non-Gaussian prior, and expect that it would learn an appropriate ICA model. However, this is not the case. In Khemakhem et al. [2020] some experiments suggest that VAEs with ICA-appropriate priors are unsuited to performing non-linear ICA. In our experiments (55) we further verify this line of inquiry and show that VAEs struggle to match their aggregate posteriors to non-Gaussian priors and thus are unable to discover independent latent sources.

Though source separation can be achieved by ‘disentangling’ methods such as the $\beta$-VAE [Higgins et al. 2017] and $\beta$-TCVAE [Chen et al. 2018], these methods require post-hoc penalisation of certain terms of the VAE objective, at times inducing improper priors (in the $\beta$-TCVAE in particular [Mathieu et al. 2019]). Further, precise tuning of this penalisation, a form of soft supervision, is key to getting appropriate representations [Rolinek et al., 2019; Locatello et al., 2019]. Stühmer et al. [2019] obtains a variety of non-linear ICA using VAEs with sets of Generalised Gaussian priors, but even then $\beta$ penalisation is required to obtain ‘disentangled’ representations.

As such there is a need for probabilistic DGMs that can separate sources without added hyperparameter tuning and that can do so by matching ICA-appropriate priors. Our solution combines linear ICA with a dimensionality-preserving invertible flow $f_\theta$. The flow acts between our data space of dimensionality and the representation fed to the linear ICA generative model; learning a representation that is well fit by the simple, linear ICA model. As we demonstrate in experiments (55), this hybrid model, which we call Bijecta, succeeds where VAEs fail: it can match non-Gaussian priors and is able to discover independent latent sources on image datasets.

3.1 A Linear ICA base distribution for flows

Our aim here is to develop a non-square ICA method that is both end-to-end differentiable and computation-efficient, such that it can be trained jointly with a flow via stochastic gradient descent. We begin by choosing our base ICA source distribution to be a set of independent generalised Gaussian distributions, Eq (1) with $\mu = 0$, $\alpha = 1$ and $\rho$ varying per experiment; and the ICA model’s likelihood to a be a Gaussian.

\[ p(s_i) = GG(s_i|\mu = 0, \alpha = 1, \rho), \text{for } i \in \{1, \ldots, d_s\}, \]
\[ p(z|s) = N(x|A s, \Sigma_\theta), \]
where $A \in \mathbb{R}^{d_s \times d_x}$ is our (unknown) ICA mixing matrix, which acts on the sources to produce a linear mixture; and $\Sigma_\theta$ is a learnt or fixed diagonal covariance. This linear mixing of sources yields an intermediate representation $z$ that is then mapped to the data by a flow. Our model has three sets of variables: the observed data $x$, the flow representation $z = f^{-1}(x)$, and ICA latent sources $s$. It can be factorised as
\[ p_\theta(x, s) = p_\theta(x|s)p(s) = p(z|s)p(s) \left| \det \frac{\partial f^{-1}}{\partial z} \right| \quad (5) \]

While it is simple to train a flow by maximum likelihood method when we have a simple base distribution in $Z$, here to obtain a maximum likelihood objective we would have to marginalise out $s$ to obtain the evidence in $Z$; a computationally intractable procedure:
\[ p(z; A, \Sigma_\theta) = \int ds \ p(z|s; A, \Sigma_\theta)p(s). \quad (6) \]
A contemporary approach is to use amortised variational inference for the linear ICA part of our model. This means we introduce an approximate amortised posterior for $s$ and perform importance sampling on
Learning Bijective Feature Maps for Linear ICA

Figure 2: The generative model (a) and variational posterior (b), as defined in Eq (8).

Eq (8), taking gradients through our samples using the reparameterisation trick [Kingma & Welling, 2014; Rezende et al., 2014]. Amortised stochastic variational inference offers numerous benefits: it scales training to large datasets by using stochastic gradient descent, our trained model can be applied to new data with a simple forward pass, and we are free to choose the functional & probabilistic form of our approximate posterior. Further, our ICA model is end-to-end differentiable, making it optimal for jointly training with a flow.

We choose a linear mapping in our posterior, with $q_\phi(s|z) = \text{Laplace}(s|A^Tz, b_\phi)$, where we have introduced variational parameters $\phi = \{A^+, b_\phi\}$ corresponding to an unmixing matrix and a diagonal divergence. Using samples from this posterior we can define a lower bound $\mathcal{L}$ on the evidence in $\mathcal{Z}$

$$\log p(z; A, \Sigma_\theta) \geq \mathcal{L}(z; \phi, A, \Sigma_\theta)$$

Using the change of variables equation, Eq (7), and the lower bound on the evidence for ICA in Eq (7) for $\mathcal{Z}$, we can obtain a variational lower bound on the evidence for our data $x$ as the sum of the ICA model’s ELBO (acting on $z$) and the log determinant of the flow:

$$\log p_\theta(x; A, \Sigma_\theta) \geq \mathcal{L}(x; \theta, \phi, A, \Sigma_\theta)$$

As such our model is akin to a flow model, but with an additional latent variable $s$; the base distribution $p(z)$ of the flow is defined through marginalizing out the linear mixing of the sources. We refer to a model with $n$ non-linear splines mapping from $\mathcal{X}$ to $\mathcal{Z}$ as an $n$-layer Bijecta model.

In the case of non-square ICA, where our ICA model is not perfectly invertible, errors when reconstructing a mapping from $\mathcal{S}$ to $\mathcal{Z}$ may amplify when mapping back to $\mathcal{X}$. To mitigate this we add an additional regularisation term in our loss that penalises the $L_1$ error of each point when reconstructed into $\mathcal{X}$. This penalisation can be weighted according to the importance of high-fidelity reconstructions for a given application.

We attempted to train Bijecta with unconstrained mixing and unmixing matrices, but found that jointly training a linear model with a powerful flow was not trivial and models failed to converge when naively optimising Eq (8). We found it crucial to appropriately constrain the unmixing matrix to get models to converge. We detail these constraints in the next section.

4 Whitening in $A^+$, without SVD

What are good choices for the mixing and unmixing matrices? Recall in Sec 2.2 we discussed various traditional approaches to constraining the unmixing matrix. For our flow-based model, design choices as to the parameterisation of $A^+$ stabilise and accelerate training. As before, the mixing matrix $A$ is unconstrained during optimisation. However, without the constraints on $A^+$ we describe in this section, we found that joint training of a flow with linear ICA did not converge.

Recall Eq (2) — linear ICA methods carry out whitening $W$, performing dimensionality reduction projecting from a $d_z$-dimensional space to a $d_z$-dimensional space, and the remaining rotation and scaling operations are square. When training with a flow the powerful splines we are learning can fulfill the role of the square matrices $R$ and $\Phi$, but doing this ahead of the whitening itself. Put another way, the outputs from the flow can be learnt such that they are simply a whitening operation away from being effective ICA representations in $\mathcal{S}$. Thus, to minimise the complexities of jointly training a powerful flow with a small linear model, we can simply set $A^+ = W$, such that the unmixing matrix projects from $d_z$ to $d_z$ and is decorrelating. Statistical independence will come from the presence of the KL term in Eq (7); the flow will learn to give $z$ representations that, when whitened, are good ICA representations in $\mathcal{S}$. See Fig 4 for a visual illustration of this process and a comparison with the steps involved in linear ICA.

In previous linear ICA methods, the whitening procedure $W$ has been derived in some data-aware way. A common choice is to whiten via the Singular Value Decomposition (SVD) of the data matrix, where $W = \Sigma U^T$, $\Sigma$ is the rectangular diagonal matrix of singular values of $X$, and the columns of $U$ are the left-singular vectors. Computing the SVD of the whole dataset is expensive for large datasets; for us, in the context of Bijecta, we would be re-calculating the SVD of the representations $Z = f^{-1}(X)$ of the entire dataset after every training step. One route around this would be an online calculation of the whitening matrix [Cardoso & Laheld, 1996; Hyvärinen et al., 2001]. This introduces an extra optimisation process that also has to be tuned, and would interact with the training of the flow.

To tackle these shortcomings of existing whitening methods, we propose a new method for linear non-square ICA that uses Johnson–Lindenstrauss (JL) transforms (also known as sketching) [Woodruff, 2014].
which not only works effectively as a linear ICA method, but also works in conjunction with a flow model. These JL transforms have favourable properties for ICA, as we demonstrate in theoretical results. Further, this method samples part of the whitening matrix at initialisation and leaves it fixed for the remainder of training, requiring no hyper-parameter tuning and making it extremely computationally efficient. This method is novel and efficient when used as a whitening method within linear ICA, and when combined with a flow as in Bijecta is a powerful method for non-linear ICA as we demonstrate in experiments.

4.1 Approximately-Stiefel matrices

We have set \( A^+ = W \), the whitening matrix. \( W \) has two aims in non-square ICA. The first is dimensionality reduction, projecting from a \( d_x \)-dimensional space to a \( d_s \)-dimensional space. The second is to decorrelate the data it transforms, meaning that the resulting projection will have unit variance and mutually uncorrelated components. More formally we wish for \( W \) of dimensionality \( d_s \times d_x \) to be decorrelating.

The set of orthogonal decorrelating rectangular matrices lie on the Stiefel Manifold \( \mathbb{S}^{r \times c} \) denoted \( \mathcal{V} \). For matrices with \( r \) rows and \( c \) columns, a matrix \( G \in \mathcal{V}(r,c) \) iff \( GG^* = I \) (\( G^* \) the conjugate transpose of \( G \)). Constraining the optimisation of \( W \) to this manifold can be computationally expensive and complex (Bakir et al. 2004; Harandi & Fernando 2016; Siegel 2019) and instead we choose for \( W \) to be approximately Stiefel, that is to lie close to \( \mathcal{V}(d_s, d_x) \). This is justified by the following theorem, proved in Appendix A:

**Theorem 1.** Let \( G \) be a rectangular matrix and \( \tilde{G} \) be its projection onto \( \mathcal{V}(r,c) \). As the Frobenius norm \( \|G - \tilde{G}\| \to 0 \) we have that \( \|GXX^TGT - \Psi\| \to 0 \), where \( GXX^TGT \) is the cross-correlation of the projection of data \( X \) by \( G \), and \( \Psi \) is some diagonal matrix.

Simply put, this shows that as a matrix \( G \) approaches the Stiefel manifold \( \mathcal{V}(r,c) \) the off-diagonal elements of the cross-correlation matrix of the projection \( GX \) are ever smaller, so \( G \) is ever more decorrelating. Given these properties we want our whitening matrix to lie close to the Stiefel manifold.

4.1.1 Johnson-Lindenstrauss projections

By Theorem 1 we know that we want our whitening matrix to be close to \( \mathcal{V}(d_s, d_x) \). How might we enforce this closeness? By the definition of the Stiefel manifold, we can intuit that a matrix \( G \) will lie close to this manifold if \( GG^T \approx I \). We formalise this as:

**Theorem 2.** Let \( G \in \mathbb{R}^{d_x \times d_s} \) and let \( \tilde{G} \) be its projection onto \( \mathcal{V}(d_s, d_x) \). As the Frobenius norm \( \|GG^T - I\| \to 0 \), we also have \( \|G - \tilde{G}\| \to 0 \).

The proof for this is presented in Appendix D. Using this theorem, we now propose an alternative to SVD-based whitening. Instead of having \( W = \Sigma^{-1}U^T \) be the result of SVD on the data matrix, we define our whitening matrix as a data-independent Johnson–Lindenstrauss transform. We must ensure that \( W \), our rectangular matrix, is approximately orthogonal, lying close to the manifold \( \mathcal{V}(d_s, d_x) \). More formally by Theorem 2 our goal is to construct a rectangular matrix \( W \) such that \( WW^T \approx I \).

We construct approximately orthogonal matrices for \( W \) by way of Johnson-Lindenstrauss (JL) Projections (Johnson & Lindenstrauss 1984). A JL projection \( W \) for \( \mathbb{R}^{d_x} \to \mathbb{R}^{d_s} \) is sampled at initialisation from a simple binary distribution Achlioptas (2003):

\[
W_{i,j} = \begin{cases} 
+1/\sqrt{d_x}, & \text{with probability } \frac{1}{2} \\
-1/\sqrt{d_x}, & \text{with probability } \frac{1}{2}
\end{cases}
\]

This distribution satisfies \( \mathbb{E}[WW^T] = I \), and such a draw has \( WW^T \approx I \). We choose to fix \( W \) after initialisation such that \( A^+ = W \) never updates, greatly simplifying optimisation.
Learning Bijective Feature Maps for Linear ICA

Figure 4: Here we run linear ICA on a pair of images (a) that are mixed linearly (mix = w1 * image1 + w2 * image2) (b) to form a dataset with 512 points. In both cases w1 and w2 are sampled from a uniform distribution. We plot the mixing matrix A for our JL-Cayley model with a quasi-uniform GG prior with \( \rho = 10 \) (c) and for FastICA (Hyvärinen & Oja, 1997) as a benchmark. A should recover the source images, which occurs for both models.

Figure 5: Here we demonstrate that Bijecta is capable of unmixing non-linearly mixed sources, better than VAEs with ICA-appropriate priors. We take a dSprites heart and, using a randomly sampled affine transformation, move it around a 32 by 32 background (a). With 2-D GG priors with \( \rho = 10 \) for a convolutional VAE (b) and for Bijecta (c) we plot the generations resulting from traversing the 2-D latent-source space in a square around the origin. We sketch the learnt axis of movement of the sprite with white lines. In (b) the VAE does not ascribe consistent meaning to its latent dimensions. It has failed to discover consistent independent latent sources: it has a sudden change in the learnt axes of movement along the second dimension, as seen by the kink in the white vertical lines. In (c) Bijecta is able to learn a simple affine transformation along each latent dimension, consistently spanning the space. In Fig B.2 we show the posterior distributions of both these models and show that Bijecta is better able to match the GG prior than the VAE, supporting our findings here.

5 Experiments

Here we show that our approach outperforms VAEs and flows with ICA-priors at discovering ICA sources in image data. But first, as a sanity check, we show that a linear ICA model using JL projections to whiten can successfully unmix linearly mixed sources in Fig 4. For details on how to implement such a linear ICA model, see Appendix E. We take a pair of images from dSprites and create linear mixtures of them. We see that linear ICA with JL projections can successfully discover the true sources, the images used to create the mixtures, in the columns of A.

Affine Data Given that we have established that our novel theory for decorrelating matrices can produce standalone linear ICA models, we now want to ascertain that our hybrid model performs well in non-linear mixing situations. To do so we create a dataset consisting of a subset of dSprites where we have a light-blue heart randomly uniformly placed on a black field. The true latent sources behind these randomly sampled affine transformations are simply the coordinates of the heart. First, in Fig B.2 we demonstrate that linear ICA models are unable to uncover the true latent sources. As expected non-linear mixing regimes motivate the use of flexible non-linear models.

We now demonstrate that Bijecta can uncover the latent sources underpinning these affine transformations, whereas VAEs with ICA-appropriate priors fail to do so. For details of VAE architecture, see Appendix I. These VAEs are able to learn to reconstruct data well, but the learnt latent space does not correspond to the underlying statistically independent sources (see Figs 1 and B.2). In fact for VAEs the effect of the latent variables is not consistent throughout the latent space, as seen in Fig 5. For Bijecta, the learnt latent space corresponds to the underlying statistically independent sources (see Figs 1 and B.2), and the meaning of the latent variables is consistent in Fig 5. Further in Fig 5 the model seems able to extrapolate outside the training domain: it generates images where the heart is partially rendered at the edges of the frame, even removing the heart entirely at times, even though such images are not in the training set.
An ersatz test of this can be done by synthesising im-
VAEs in learning factorisable representations.
where
To quantify this source-separation numerically, we mea-
the flow has not discovered axis-aligned transforms, whereas Bijecta has learned informative latent dimensions:
variables due to dependence (Watanabe, 1960). It di-

tation of the distribution:
son & Roberts, 2001). Formally, it is the
rect learnt the generating sources
lerative to flow models with heavy-tailed base distribu-
our model’s ability to automatically uncover sources relative to flow models with heavy-tailed base distributions. We do so by measuring the cumulative explained variance by the dimensions in \( Z \) for both models. If a small number of dimensions explains most of \( Z \)’s variance then the model has learnt a bijection which only requires a small number of dimensions to be inver-
tible. It has in effect learnt the generating sources underpinning the data.
In Fig 7 we show that Bijecta induces better-

**Natural Images** The previous experiments show that our model is capable of isolating independent sources on toy data. We complement this finding with experiments on a more complex natural image dataset, CelebA, and show that here too our model outperforms VAEs in learning factorisable representations.

An ersatz test of this can be done by synthesising images where we sample from a factorised approximation of Bijecta’s posterior. If the learned latent sources are actually independent, then the posterior over latent sources given the entire dataset should factorize into a product across dimensions, i.e. \( q(s) = \prod_i q(s_i) \).
In this case, we can fit an approximation to the posterior by fitting \( d_s \) independent one-dimensional density estimates on \( q(s_i) \). If the sources are not independent, then this factorized approximation to the posterior will be missing important correlations and dependencies. In Fig 6a samples from this factorised approximation look reasonable, suggesting that Bijecta has learnt representations that are statistically independent.

To quantify this source-separation numerically, we measure the total correlation (TC) of the aggregate posterior of Bijecta \( (q(s|z)) \) and VAEs \( (q(z|x)) \) as [Chen et al., 2018] do. Intuitively, the TC measures how well a distribution is approximated by the product of its marginals – or how much information is shared between variables due to dependence [Watanabe, 1960]. It directly measures how well an ICA model has learnt decorrelated and independent latent representations [Everson & Roberts, 2001]. Formally, it is the KL divergence between a distribution \( r(\cdot) \) and a factorised representation of the distribution: 

\[
\text{TC} = \text{KL}(r(s)|| \prod_i r(s_i)),
\]

where \( i \) indexes over (latent) dimensions.

In Table 1 we show that Bijecta learns an aggregate posterior with significantly lower TC values than both VAEs with Laplace priors, and \( \beta \)-TCVAEs – which in their training objective penalise the TC by a factor \( \beta \) (Chen et al., 2018). Our model has learnt a better ICA solution. We also include numerical results in Appendix J.1 showing that Bijecta outperforms linear ICA on a variety of natural image datasets.

<table>
<thead>
<tr>
<th></th>
<th>Laplace-VAE</th>
<th>( \beta )-TCVAE</th>
<th>Laplace-Bijecta</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC</td>
<td>106.7 ( \pm ) 0.9</td>
<td>55.7 ( \pm ) 0.1</td>
<td>\textbf{13.1 ( \pm ) 0.4}</td>
</tr>
</tbody>
</table>

**Dimensionality reduction on flow models** To conclude, having shown that Bijecta outperforms VAEs on a variety of non-linear ICA tasks, we now contrast our model’s ability to automatically uncover sources relative to flow models with heavy-tailed base distributions. We do so by measuring the cumulative explained variance by the dimensions in \( Z \) for both models. If a small number of dimensions explains most of \( Z \)’s variance then the model has learnt a bijection which only requires a small number of dimensions to be invertible. It has in effect learnt the generating sources underpinning the data.

In Fig 7 we show that Bijecta induces better-
Learning Bijective Feature Maps for Linear ICA

Figure 7: Explained variance plots for the embedding in $Z$, as measured by the sums of the eigenvalues of the covariance matrix of the embeddings, for both our Bijecta model and for an RQS model of equivalent size trained with a Laplace base distribution ($GG$ distribution with $\rho = 1$). For both Fashion-MNIST (left) and CIFAR 10 (right) datasets we see that the Bijecta model has learned a compressive flow, where most of the variance can be explained by only a few linear projections. The shaded region denotes the first 64 dimensions, corresponding to the size of the target source embedding $S$.

compressed representations in $Z$ than non-compressive flows. We plot the eigenvalues of the covariance matrix on the output of the flow, i.e. on $\text{Cov}(f(X))$, to see how much of the total variance in the learned feature space $Z$ can be explained in a few dimensions. In doing so we see that a flow trained jointly with a linear ICA model with $d_s = 64$ effectively concentrates variation into a small number of intrinsic dimensions; this is in stark contrast with the RQS flows trained with only a Laplace base distribution. This demonstrates that our model is able to automatically detect relevant directions on a low dimensional manifold in $Z$, and that the bijective component of our model is better able to isolate latent sources than a standard flow.

For a visual illustration of this source separation we show the difference in generated images resulting from smoothly varying along each dimension in $S$ for Bijecta models and in $Z$ for flows in Fig 6. Bijecta is clearly able to discover latent sources, whereby it learns axis-aligned transformations of CelebA faces, whereas a flow with equivalent computational budget and a heavy-tailed base distribution is not able to.

All flow-based baselines are trained using the objective in Eq 4, using Real-NVP style factoring-out (Durkan et al., 2019; Dinh et al., 2015), and are matched in size and neural network architectures to the flows of Bijecta models. See Appendix I for more details.

6 Related Work

One approach to extend ICA to non-linear settings is to have a non-linear mapping acting on the independent sources and data (Burel, 1992; Deco & Brauer, 1995; Yang et al., 1998; Valpola et al., 2003). In general, non-linear ICA models have been shown to be hard to train, having problems of unidentifiability: the model has numerous local minima it can reach under its training objective, each with potentially different learnt sources (Hyvärinen & Pajunen, 1999; Karhunen, 2001; Almeida, 2003; Hyvarinen et al., 2019). Some non-linear ICA models have been specified with additional structure to reduce the space of potential solutions, such as putting priors on variables (Lappalainen & Honkela, 2000) or specifying the precise non-linear functions involved (Lee & Koehler, 1997; Taleb, 2002). Recent work shows that conditioning the source distributions on some always-observed side information, say time index, can be sufficient to induce identifiability in non-linear ICA (Khemakhem et al., 2020).

Modern flows were first proposed as an approach to non-linear square ICA (Dinh et al., 2015), but are also motivated by desires for more expressive priors and posteriors (Kingma et al., 2016; Papamakarios et al., 2019). Early approaches, known as symplectic maps (Deco & Brauer, 1995; Parra et al., 1995, 1996), were also proposed for use with ICA. Flows offer expressive dimensionality-preserving (and sometimes volume-preserving) bijective mappings (Dinh et al., 2017; Kingma & Dhariwal, 2018). Flows have been used to provide feature extraction for linear discriminative models (Nalisnick et al., 2019). Orthogonal transforms have been used in normalizing flows before, to improve the optimisation properties of Sylvester flows (Van Den Berg et al., 2018; Golinski et al., 2019). Researchers have also looked at constraining neural network weights to the Stiefel-manifold (Li et al., 2020).

7 Conclusion

We have developed a method for performing non-linear ICA large high-dimensional image datasets which combines state-of-the-art flow-based models and a novel theoretically grounded linear ICA method. This model succeeds where existing probabilistic deep generative models fail: its constituent flow is able to learn a representation, lying in a low dimensional manifold in $Z$, under which sources are separable by linear unmixing. In source space $S$, this model learns a low dimensional, explanatory set of statistically independent latent sources.
Acknowledgments

This research was directly funded by the Alan Turing Institute under Engineering and Physical Sciences Research Council (EPSRC) grant EP/N510129/1. AC was supported by an EPSRC Studentship. MW was supported by EPSRC grant EP/G03706X/1. CH was supported by the Medical Research Council, the Engineering Physical Sciences Research Council, Health Data Research UK, and the Li Ka Shing Foundation SR gratefully acknowledges support from the UK Royal Academy of Engineering and the Oxford-Man Institute.

We thank Tomas Lazauskas, Jim Madge and Oscar Giles from the Alan Turing Institute’s Research Engineering team for their help and support.

References


