Binary Independent Component Analysis via Non-stationarity

Abstract

We consider independent component analysis of binary data. While fundamental in practice, this case has been much less developed than ICA for continuous data. We start by assuming a linear mixing model in a continuous-valued latent space, followed by a binary observation model. Importantly, we assume that the sources are non-stationary; this is necessary since any non-Gaussianity would essentially be destroyed by the binarization. Interestingly, the model allows for closed-form likelihood by employing the cumulative distribution function of the multivariate Gaussian distribution. In stark contrast to the continuous-valued case, we prove non-identifiability of the model with few observed variables; our empirical results imply identifiability when the number of observed variables is higher. We present a practical method for binary ICA that uses only pairwise marginals, which are faster to compute than the full multivariate likelihood. Experiments give insight into the requirements for the number of observed variables, segments and latent sources, that allow for the model to be estimated.

1 INTRODUCTION

Despite significant progress in both linear and nonlinear ICA in recent years [Hyvärinen and Morioka, 2016, Hyvärinen et al., 2019, Khemakhem et al., 2020], ICA for binary data remains a challenging and important problem as binary data is abundant in various fields, such as bioinformatics, health informatics, social sciences, natural language, and electrical engineering. An ICA model for binary data may also open new opportunities in solving problems closely related to ICA, such as causal discovery [Shimizu et al., 2006] and feature extraction [Hyvärinen and Morioka, 2016].

Methods for binary ICA have been proposed based on either binary or continuous-valued independent components. In the case of binary components, Himberg and Hyvärinen [2001], Nguyen and Zheng [2011] assumed an OR mixture model. Some extensions of Latent Dirichlet can be seen as binary ICA [Podosinnikova et al., 2015, Buntine and Jakulin, 2005]. On the other hand, Kabán and Bingham [2006] presented an approach based on a latent linear model and binarized observations, although the components were restricted to the unit interval, which restricts its applicability. Recently, Khemakhem et al. [2020] presented a nonlinear ICA model (iVAE) that can employ binarized observations, making several contributions that we can build on.

Our goal here is to study the prospects of ICA for binary data using a model, that is theoretically analyzable and intuitively appealing. Identifiability of such a model is crucial to investigate, and we also want a consistent estimator which is not based on approximations whose validity is not clear. None of the approaches above fulfills all these criteria.[1]

We propose a binary ICA model inspired by recent developments in nonlinear ICA. We formulate a latent linear model with a separate binarizing measurement equation. Crucially, we assume the components to be non-stationary, which is a powerful principle and very useful here because any non-Gaussianity (commonly employed in ICA) may be destroyed by binarization. Thus, we obtain a binary ICA model whose likelihood can actually be described in closed form via the multivariate Gaussian cumulative distribution function. We further propose to combine the likelihood with a moment-matching approach to obtain a fast and accurate estimation algorithm. In fact, due to the model structure, pairwise marginal distributions of non-binarized data can be accurately estimated from binary sample data and the likelihood can be computed directly from them. We investigate the identifiability of the model, and somewhat surprisingly, we show that low-dimensional models are in

[1] As noted in the Corrigendum of Khemakhem et al. [2020] (v4 on arxiv), their initial identifiability proof for a discrete non-linear ICA model is incorrect.

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

fact non-identifiable—while higher-dimensional models are (empirically) shown to be identifiable.

2 A MODEL FOR BINARY ICA

We define here a binary counterpart of the linear ICA model. In particular, we consider here a model with non-stationarity: we assume the data consists of n observed variables, divided into n_u segments which express the non-stationarity. Thus, each data point has a segment index u assigned to it. This additionally observed variable u assumes categorical values, which is a special case of the auxiliary variable framework of Khemakhem et al. [2020]. Such non-stationarity based on a segment-wise Gaussian model is well-known in linear ICA [Pham and Cardoso, 2001, Miettinen et al., 2017]. It is not only natural in the case of non-stationary time series, but also when there is any other external discrete variable, such as the experimental condition or intervention, or even a class label which modulates the distribution of the data.

We assume the data is generated from n_z latent variables (independent components, or sources), collected into a latent random vector $\mathbf{z}^u$, which are generated independently of each other from a Gaussian distribution. Crucially, the parameters of the Gaussian distribution change as a function of the segment as

$$\mathbf{z}^u \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}}^u, \boldsymbol{\Sigma}_{\mathbf{z}}^u)$$

where $\boldsymbol{\Sigma}_{\mathbf{z}}^u$ is a diagonal matrix of the source variances in segment u.

We define "intermediate" variables $\mathbf{y}^u$ which are a linear mixing of the sources by a mixing matrix $\mathbf{A}$ with n rows and n_z linearly independent columns

$$\mathbf{y}^u = \mathbf{A}\mathbf{z}^u \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu}_{\mathbf{z}}^u,\ \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{z}}^u\mathbf{A}^\intercal) \tag{1}$$

Note that the mixing matrix $\mathbf{A}$ is constant over the segments u [Pham and Cardoso, 2001].

While some work in ICA considers noisy continuous observations by adding noise to $\mathbf{y}^u$, we can consider here binarized observations $\mathbf{x}^u$ instead. Binarization is done using a linking function σ so that the probability of ith element of $\mathbf{x}^u$ being 1 is:

$$P(x_i^u = 1) = \sigma(y_i^u)$$

We use a linking function based on the Gaussian CDF (cumulative distribution function):

$$\sigma(y_i^u) = \Phi(\sqrt{\frac{\pi}{8}} y_i^u | 0, 1)$$

where Φ is the cumulative distribution function of the Gaussian distribution, here with mean 0 and variance 1. We use $\sqrt{\frac{\pi}{8}}$ as the coefficient to match closely to the often employed sigmoid function $\sigma(y_i) = \frac{1}{1+e^{-y_i}}$ [Waissi and Rossin, 1996, Li, 2021]. The Gaussian CDF based linking function has certain nice algebraic properties compared to the sigmoid, as we will see in Section 3.

Furthermore, the linking function has the following intuitive interpretation. Take y_i^u, add independent noise ϵ form $\mathcal{N}(0, \frac{8}{\pi})$, and binarize y_i^u simply by a hard threshold 0 to get x_i^u. This gives the same distribution for x_i^u, since the probabilities match:

$$\begin{aligned} &P(x_i^u = 1) = P(y_i^u + \epsilon > 0) = P(\epsilon > -y_i^y) \\ &= \int_{-y_i^u}^{\infty} \mathcal{N}\left(\epsilon | 0, \frac{8}{\pi}\right) d\epsilon = \Phi\left(\sqrt{\frac{\pi}{8}} y_i^u | 0, 1\right). \end{aligned}$$

A binary ICA model $\mathcal{M} = (\mathbf{A}, \{\boldsymbol{\mu}_{\mathbf{z}}^u\}_u, \{\boldsymbol{\Sigma}_{\mathbf{z}}^u\}_u)$ thus consists of the following parameters: mixing matrix $\mathbf{A}$, the means $\boldsymbol{\mu}_{\mathbf{z}}^u$ and the diagonal (co)variance matrices $\boldsymbol{\Sigma}_{\mathbf{z}}^u$ for all segments u, denoted by $\{\boldsymbol{\mu}_{\mathbf{z}}^u\}_u$ and $\{\boldsymbol{\Sigma}_{\mathbf{z}}^u\}_u$. Consequently it defines a distribution for binary vector $\mathbf{x}^u$ in each segment indexed by u.

3 THE LIKELIHOOD

A surprising observation regarding the the latent variable model defined in Section 2, is that we can calculate the likelihood in closed-form by employing the multivariate Gaussian cumulative distribution function. For example, the model defines the probability of the data vector of all ones, denoted by $\mathbf{1}$, as:

$$\begin{aligned} P(\mathbf{x}^u = \mathbf{1}|\mathcal{M}) &= \int P(\mathbf{x}^u = \mathbf{1}|\mathbf{y}^y) P(\mathbf{y}^u|\mathcal{M}) d\mathbf{y} \\ &= \int \Phi\left(\sqrt{\frac{\pi}{8}} \mathbf{y}^u | \mathbf{0}, \mathbf{I}\right) \mathcal{N}(\mathbf{y}^u | \mathbf{A}\boldsymbol{\mu}_{\mathbf{z}}^u, \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{z}}^u\mathbf{A}^\intercal) d\mathbf{y} \end{aligned}$$

where the univariate Gaussian cumulative distribution functions are written as a multivariate Gaussian cumulative distribution function Φ with an identity covariance matrix. The benefit of using a Gaussian cumulative distribution function based linking function comes into play here, as the value of the integral is directly a value of a multivariate Gaussian cumulative distribution function [Waissi and Rossin, 1996, Li, 2021]: The above formula actually specifies the probability of first drawing $\mathbf{y}^u$, multiplying it with $\sqrt{\pi/8}$ and then, independently, drawing standard Gaussian variable $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ that is elementwise smaller. We therefore have:

$$P(\mathbf{x}^u = \mathbf{1}|\mathcal{M}) = P\left(\mathbf{n} - \sqrt{\frac{\pi}{8}} \mathbf{y}^u < \mathbf{0}\right)$$

This motivates us to define a random vector $\mathbf{q}^u$, an important construct in the following developments as:

$$\mathbf{q}^u \quad = \quad \mathbf{n} - \sqrt{\frac{\pi}{8}} \mathbf{y}^u, \tag{2}$$

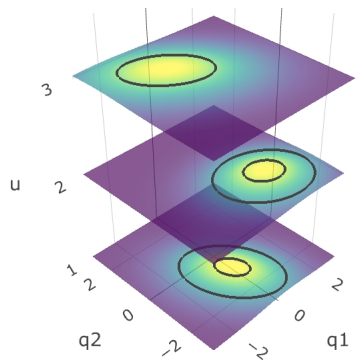

Figure 1: Binary ICA model for two observed variables and three segments. For each segment there is a bivariate Gaussian distribution on $\mathbf{q}^u$, the probability of an assignment to the binary observed variables is the probability mass in the corresponding quadrant.

which is simply a noisy, rescaled and sign-flipped version of the linear mixture $\mathbf{y}^u$. In fact, since $\mathbf{q}^u$ is the sum of two independent Gaussian random vectors it also has a Gaussian distribution $\mathbf{q}^u \sim \mathcal{N}\left(\boldsymbol{\mu}^u_{\mathbf{q}}, \boldsymbol{\Sigma}^u_{\mathbf{q}}\right)$ with:

$$\boldsymbol{\mu}^u_{\mathbf{q}} = -\sqrt{\frac{\pi}{8}}\mathbf{A}\boldsymbol{\mu}^u_{\mathbf{z}}, \quad (3)$$

$$\boldsymbol{\Sigma}^u_{\mathbf{q}} = \mathbf{I} + \frac{\pi}{8}\mathbf{A}\boldsymbol{\Sigma}^u_{\mathbf{z}}\mathbf{A}^\intercal. \quad (4)$$

The probability of data vector of ones in segment u is then:

$$P(\mathbf{x}^u = \mathbf{1}|\mathcal{M}) = P\left(\mathbf{q}^u < \mathbf{0}\right) = \Phi\left(\mathbf{0}|\boldsymbol{\mu}^u_{\mathbf{q}}, \boldsymbol{\Sigma}^u_{\mathbf{q}}\right), \quad (5)$$

where the cumulative distribution function of the multivariate Gaussian Φ has all variables integrated from $-\infty$ to 0; it is readily implemented in basic packages [Genz and Bretz, 2009].

Similar derivation gives the probabilities for other assignments to $\mathbf{x}^u$. These probabilities can be expressed compactly for all value assignments as:

$$P(\mathbf{x}^u|\mathcal{M}) = \Phi\left(l(\mathbf{x}^u), u(\mathbf{x}^u)|\boldsymbol{\mu}^u_{\mathbf{q}}, \boldsymbol{\Sigma}^u_{\mathbf{q}}\right) \quad (6)$$

in which the multivariate Gaussian probability density function is integrated from the lower bound $l(\mathbf{x}^u)$ to the upper bound $u(\mathbf{x}^u)$, with the ith elements in the bounds defined by:

$$l(\mathbf{x}^u)[i] = \begin{cases} -\infty \text{ if } x^u_i = 1 \\ 0 \text{ otherwise} \end{cases} \quad u(\mathbf{x}^u)[i] = \begin{cases} 0 \text{ if } x^u_i = 1 \\ \infty \text{ otherwise} \end{cases}$$

Importantly, this formulation allows for a particularly clear intuitive interpretation of the model. Figure 1 shows this for two observed variables and three segments. For each segment the model defines a bivariate Gaussian distribution for $\mathbf{q}^u$, depicted by colors and contours on the planes. The probability for an assignment of the observed binary variables $\mathbf{x}^u$ in a segment is simply the probability mass in a corresponding quadrant. The multivariate Gaussian distributions for $\mathbf{q}^u$ in each segment are related in the sense that they are formed by the same mixing matrix performing on independent sources particular to the segment.

The log-likelihood of the whole data set can then be calculated as

$$l = \sum_u \sum_{\mathbf{x}^u} c(\mathbf{x}^u) \log \Phi(l(\mathbf{x}^u), u(\mathbf{x}^u)|\boldsymbol{\mu}^u_{\mathbf{q}}, \boldsymbol{\Sigma}^u_{\mathbf{q}}), \quad (7)$$

where $c(\mathbf{x}^u)$ is the count of the data points $\mathbf{x}^u$ in segment u and the sum is taken over all assignments to $\mathbf{x}^u$ and u.

4 ON IDENTIFIABILITY

Many ICA models can only be identified up to scaling and permutation indeterminacies of the sources [Hyvärinen et al., 2001, Khemakhem et al., 2020]. Straightforwardly we can see that those limitations apply for our model as well. By re-ordering columns of the mixing matrix and the sources, the implied distribution is unaffected; similarly, we can counteract the scaling (or sign-flip) of the mixing matrix columns by scaling (or sign-flipping) the sources. However, binarization actually induces additional indeterminacies as we will show next.

4.1 THE BINARIZATION INDETERMINACY

Recall that the probability of an assignment to binary $\mathbf{x}^u$ is given by the probability of the Gaussian $\mathbf{q}^u$ landing in different regions in Equation 5. But note that the probability in Equation 5 stays exactly the same even if $\mathbf{q}^u$ is multiplied by a diagonal matrix $\mathbf{Q}^u$, possibly different for each segment u, with positive entries on the diagonal:

$$P\left(\mathbf{q}^u < \mathbf{0}\right) = P\left(\mathbf{Q}^u\mathbf{q}^u < \mathbf{0}\right).$$

This is valid even if the elementwise operator is $>$ or a mixture of $>$ and $<$.[2] Figure 2 shows an example of this equivalence relation for one segment and two observed variables. The two Gaussian distributions for $\mathbf{q}^u$ represented by the blue and red contours imply the exact same joint distribution for binary observed variables $\mathbf{x}^u$. The amount of mass in each of the 4 quadrants is exactly the same. This means that we essentially lose all scale information on $\mathbf{q}^u$ in the binarization.

Then, two binary ICA models $\mathcal{M} = (\mathbf{A}, \{\boldsymbol{\mu}^u_{\mathbf{z}}\}_u, \{\boldsymbol{\Sigma}^u_{\mathbf{z}}\}_u)$ and $\hat{\mathcal{M}} = (\hat{\mathbf{A}}, \{\hat{\boldsymbol{\mu}}^u_{\mathbf{z}}\}_u, \{\hat{\boldsymbol{\Sigma}}^u_{\mathbf{z}}\}_u)$ are indistinguishable if

[2] For the probability of $\mathbf{x}^u$ being all ones, any permutation matrix $\mathbf{Q}^u$ would similarly preserve the implied probability, but the probability of some other assignment for $\mathbf{x}$ (each of which corresponds to some mixture of $>$ and $<$) may change then.

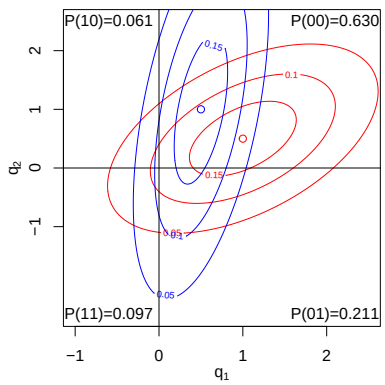

Figure 2: Two Gaussian distributions (red and blue) for a two dimensional $\mathbf{q}^u$ which imply the same binary distributions after binarization by the linking function. That is because the mass of both distributions in each of the 4 quadrants is identical.

there are positive diagonal matrices $\{\mathbf{Q}^u\}_u$ such that for each segment u means and covariances of $\mathbf{q}^u$ satisfy:

$$\hat{\boldsymbol{\mu}}^u_{\mathbf{q}} = \mathbf{Q}^u \boldsymbol{\mu}^u_{\mathbf{q}}, \quad (8)$$

$$\hat{\boldsymbol{\Sigma}}^u_{\mathbf{q}} = \mathbf{Q}^u \boldsymbol{\Sigma}^u_{\mathbf{q}} \mathbf{Q}^u, \quad (9)$$

which can be written perhaps more clearly using the model parameters (Equations 3 and 4) as:

$$\sqrt{\frac{\pi}{8}} \hat{\mathbf{A}} \hat{\boldsymbol{\mu}}^u_{\mathbf{z}} = \mathbf{Q}^u \sqrt{\frac{\pi}{8}} \mathbf{A} \boldsymbol{\mu}^u_{\mathbf{z}}, \quad (10)$$

$$\mathbf{I} + \frac{\pi}{8} \hat{\mathbf{A}} \hat{\boldsymbol{\Sigma}}^u_{\mathbf{z}} (\hat{\mathbf{A}})^\intercal = \mathbf{Q}^u (\mathbf{I} + \frac{\pi}{8} \mathbf{A} \boldsymbol{\Sigma}^u_{\mathbf{z}} \mathbf{A}^\intercal) \mathbf{Q}^u. \quad (11)$$

This limits identifiability possibilities (Section 4.2) but nevertheless also allows for the development of efficient estimation procedures in Sections 4.3 and 5.2.

4.2 THE ROW ORDER INDETERMINACY

One of consequences of the binarization indeterminacy is the following non-identifiability result concerning $n = 2$ observed variables, proven in Appendix A in the supplement.

Theorem 1. *If the row order of the 2-by-2 mixing matrix* $\mathbf{A}$ *of a binary ICA model is reversed, then the source means* $\boldsymbol{\mu}^u_{\mathbf{z}}$ *and variances* $\boldsymbol{\Sigma}^u_{\mathbf{z}}$ *can be adjusted such that the implied distributions for the observed binary* $\mathbf{x}^u$ *remain identical.*

This means that in addition to column order and scale, we have also row order indeterminacy here. Although the result may generalize to certain sparse higher dimensional models, fortunately, it does not jeopardize the estimation of higher dimensional models in general.

The result does have consequences for causal discovery [Shimizu et al., 2006, Suzuki and Inaoka, 2021, Peters et al., 2011, Inazumi et al., 2014]. Consider two structural equation models, implying opposite causal directions:

$$\mathbf{y}^u = \begin{pmatrix} 0 & 0 \\ b & 0 \end{pmatrix} \mathbf{y}^u + \mathbf{z}^u, \quad \mathbf{y}^u := \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} \mathbf{y}^u + \mathbf{z}^u.$$

where $\mathbf{z}^u$ has a Gaussian distribution in each segment u with diagonal covariance matrix $\boldsymbol{\Sigma}^u_{\mathbf{z}}$. The models correspond respectively the mixing models (compare to Equation 1):

$$\mathbf{y}^u = \begin{pmatrix} 1 & 0 \\ b & 1 \end{pmatrix} \mathbf{z}^u, \quad \mathbf{y}^u = \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} \mathbf{z}^u.$$

If we observed binarized $\mathbf{y}^u$, i.e. $\mathbf{x}^u$, we can at most identify the mixing matrix up to row order, column order and column scale. By switching the column order and then the row order of mixing matrix on the left, we get the mixing matrix on the right. Thus, unlike in the continuous case, we cannot detect the causal direction between two variables without further limiting assumptions or information on other variables.

4.3 THE CORRELATION IDENTIFIABILITY

Note that the indistinguishable models satisfying Equation 9 or Equation 11 have equal *correlation matrices* (i.e. matrices of Pearson correlation coefficients) for the random variables $\mathbf{q}^u$. The next theorem and corollary show that the correlations between elements of $\mathbf{q}^u$ are indeed theoretically identifiable from the distributions of the binary observed variables $\mathbf{x}^u$. Intuitively, the higher the correlation the more likely will the pair of binary observed variables in $\mathbf{x}^u$ receive equal assignments. The fairly technical proof is given in Appendix B in the supplement.

Theorem 2. *Two binary ICA models imply different distributions for binary observations* $\mathbf{x}^u$ *(in a given segment* u*) if the correlation matrices for* $\mathbf{q}^u$ *are not equal.*

This result is crucial for the development of our novel estimation method (Section 5.2), via the corollary:

Corollary 1. *The correlation matrix of* $\mathbf{q}^u$ *in a given segment* u *is identifiable from the distribution for binary* $\mathbf{x}^u$.

On the other hand, the following theorem recaps the well-known result [Hyvärinen et al., 2001, Pham and Cardoso, 2001] that the means do not help in estimating the mixing matrix (proven in Appendix C):

Theorem 3. *If two models* $\mathcal{M}$ *and* $\hat{\mathcal{M}}$ *with* $n = n_z$ *imply the same correlation matrices for* $\mathbf{q}^u$ *(in a given segment) then the means* $\boldsymbol{\mu}^u_{\mathbf{z}}$ *can be adjusted such that the implied binary distributions are identical.*

5 METHODS FOR BINARY ICA

Next, we present three methods for estimating the binary ICA model, building on the theory in Sections 3 and 4. The BLICA method of Section 5.2 is the main novel algorithmic contribution of the paper.

Algorithm 1 The BLICA algorithm for Binary ICA.

1: Input data recorded at n_u different segments.
2: **for** segment $u \in \{1, \ldots, n_u\}$ **do**
3: **for** each observed variable pair $\{x_i^u, x_j^u\}$ **do**
4: Estimate the correlation between q_i^u and q_j^u by maximizing the marginal pairwise likelihood of x_i^u and x_j^u (in segment u).
5: Form and regularize the correlation matrix $\mathbf{C}_\mathbf{q}^u$ obtained from the pairwise correlations.
6: Optimize scaled Gaussian likelihood with L-BFGS over sufficient statistics $\mathbf{C}_\mathbf{q}^u$ from all segments u.
7: Return the estimated mixing matrix $\mathbf{A}$ and source variances $\mathbf{\Sigma}_\mathbf{z}^u$ for all segments u.

5.1 MAXIMUM LIKELIHOOD ESTIMATION

We have already derived the likelihood of the binary ICA model in Equation 7. A straightforward approach is then to optimize this using e.g. L-BFGS [Liu and Nocedal, 1989]. The gradient involves the moments for the *truncated* multivariate Gaussian distribution, which can be obtained from R-package tmvtnorm [Wilhelm and G, 2015]. Unfortunately, the computation of the likelihood and its gradient can only be done for small models in practice, because the evaluation of multivariate Gaussian CDF is time consuming, necessitating the use of sampling-based approximations. Our experiments refer to this as full MLE.

5.2 THE BLICA METHOD

However, we can circumvent computational burden of the high dimensional Gaussian cumulative distribution function. Due to the theory in Section 4, the correlations of $\mathbf{q}^u$ convey the essential information between the binary data and the continuous mixing model. Since the marginalization properties of our model are inherited from the multivariate Gaussian, such correlations can be estimated from *pairwise* marginal distributions of elements of $\mathbf{x}^u$; in 2D the Gaussian cumulative distribution function is still quite quick to compute. Thus, we combine maximum likelihood estimation with what could be called a "moment-matching" approach as follows. We first recover the pairwise correlations of the continuous-valued $\mathbf{q}^u$ from the observed binary data on $\mathbf{x}^u$ (this is possible by Corollary 1) via MLE in 2D. Then we fit those correlations to the correlations implied by the latent linear mixing model using a more scalable MLE in the continuous-valued latent space. The resulting algorithm is summarized as Algorithm 1 and explained in detail below.

Correlation estimation. On line 4, we estimate each correlations between elements in $\mathbf{q}^u$ separately, by directly fitting the likelihood in Equation 7 in two dimensions, thus estimating $\boldsymbol{\mu}_\mathbf{q}^u$ and $\mathbf{\Sigma}_\mathbf{q}^u$. To calculate multivariate Gaussian cumulative distribution function we use the R package mvtnorm [Genz and Bretz, 2009]. We employ the GenzBretz method, which is particularly suitable for the fast evaluation needed here [Genz, 1993]. Furthermore, the estimation can be simplified [Lee and Sompolinsky, 1999]. Due to Equation 11 the diagonal of $\mathbf{\Sigma}_\mathbf{q}^u$ can be set to 1s in this step. Furthermore, since the marginal of x_i^u is

$$P(x_i^u = 1) \quad = \quad \Phi(-\boldsymbol{\mu}_\mathbf{q}^u[i]/\sqrt{\mathbf{\Sigma}_\mathbf{q}^u[i,i]}|0,1), \quad (12)$$

thus, both means in $\boldsymbol{\mu}_\mathbf{q}^u$ can be computed from the respective marginals using the 1D inverse Gaussian CDF separately [Genz and Bretz, 2009]. The univarite optimization problem for the remaining parameter in the interval $[-1, 1]$ can then be solved efficiently using a line search method [Brent, 2013]. The scalability of Algorithm 1 depends crucially on this step as $n_u \cdot (n^2 - n)/2$ correlations need to be estimated. The separately estimated correlations are collected to n_u segmentwise n-by-n correlation matrices denoted by $\mathbf{C}_\mathbf{q}^u$.

Regularization. When estimating the correlations of $\mathbf{q}^u$ from sample data, it can happen that a correlation matrix $\mathbf{C}_\mathbf{q}^u$ is close to singular or not positive definite. We use the following regularization on line 5 based a parameter r [Warton, 2008], which marks the approximate condition number targeted. The regularized correlation matrix is then

$$\frac{1}{1+\delta}(\mathbf{C}_\mathbf{q}^u + \delta\mathbf{I}), \text{ where } \delta = \max(0, \frac{\lambda_1 - r \cdot \lambda_n}{r-1}),$$

where λ_1 is the largest and λ_n the smallest eigenvalue of $\mathbf{C}_\mathbf{q}^u$. This regularization keeps the unit diagonal.

Moment Matching. Finally on line 6, we fit the model parameters (including stationary $\mathbf{A}$) to the estimated correlations $\mathbf{C}_\mathbf{q}^u$ using a Gaussian likelihood model over the different segments u (Section 2). But in contrast to the usual case where we have covariance matrices, we need to here account for the "binarization indeterminacy", resulting in additional nuisance scaling parameters, as pointed out above. We use the term scaled Gaussian likelihood to refer to ordinary multivariate Gaussian likelihood where we include additional parameters $\mathbf{Q}^u$ as the scaling factors. The fitting is thus done by the following scaled Gaussian likelihood based on the sufficient statistics $\mathbf{C}_\mathbf{q}^u$:

$$l = \sum_{u=1}^{n_u} \frac{N}{2}[-\log(\det(\mathbf{\Sigma}_\mathbf{q}^u)) - \mathrm{TRACE}(\mathbf{C}_\mathbf{q}^u(\mathbf{\Sigma}_\mathbf{q}^u)^{-1})]$$

where recall that $\mathbf{\Sigma}_\mathbf{q}^u = \mathbf{Q}^u(\mathbf{I} + \mathbf{A}\mathbf{\Sigma}_\mathbf{z}^u\mathbf{A}^T)\mathbf{Q}^u$ by Equation 11 is a function of the mixing matrix $\mathbf{A}$, source variances $\{\mathbf{\Sigma}_\mathbf{z}^u\}_u$ (diagonal, positive elements) and scaling factors $\{\mathbf{Q}^u\}_u$ (diagonal, positive elements). Note that without the scaling factors $\{\mathbf{Q}^u\}_u$, the mixing matrix $\mathbf{A}$ could be found via joint diagonalization [Miettinen et al., 2017]. Note also that due to Theorem 3 the source means need not be estimated. Here, instead we perform the fitting by maximizing this likelihood using L-BFGS [Liu and Nocedal, 1989] with respect to the aforementioned parameters.

5.3 BINARY ICA THROUGH LINEAR IVAE

Khemakhem et al. [2020] presented the identifiable Variational Autoencoder (iVAE), an approach for nonlinear ICA employing variational autoencoders [Kingma and Welling, 2014, Rezende et al., 2014] that assumes access to an additionally observed variable such that the sources are independent given the auxiliary variable; further, each source follows an exponential family distribution given the auxiliary variable. Here, we apply iVAE to estimate the binary ICA model from Section 2. As proposed by Kingma and Welling [2014] and Khemakhem et al. [2020], we use the factorized Bernoulli observational model and apply a sigmoid function element-wise to the output of the decoder to obtain the binary probability distributions. Due to the linearity of our mixing model and the segment-wise structure, we can simplify the encoder (posterior approximation) of the VAE, and make all the transformations in the iVAE affine or linear, thus greatly simplifying the system. The linear iVAE is presented in more detail in Appendix E.

5.4 ESTIMATION OF THE SOURCES

After estimating the mixing matrix $\mathbf{A}$, it may be desired to estimate the sources $\mathbf{z}^u$ as well. In the case of binary data, the individual source values cannot be accurately estimated (even up to scale and order indeterminacies) due to the inherent noise introduced by the binarization procedure. Presumably, though, if the number of observed variables is large and the number of sources is small, the estimation may be reasonable. In any case, the posterior $P(\mathbf{z}^u|\mathbf{x}^u)$ can be easily calculated after estimating the mixing matrix.

6 EXPERIMENTS

We implemented our proposed methods and baselines using R (BLICA, full MLE) and python (linear iVAE). The methods will be published online in open source. Here we investigate the identifiability of the model, as well as the finite-sample estimation performance and the scalability of our methods, also comparing to previous approaches.

Data. Data was generated from the Binary ICA model (Section 2) in the following way. Means were drawn from $\mathrm{unif}(-0.5, 0.5)$, standard deviations from $\mathrm{unif}(0.5, 3)$. Mixing matrix elements were from $\mathrm{unif}(-3, 3)$ while ensuring invertibility by resampling until the condition number (kappa) was below 20 for $n < 20$, or for $n \leq 20$ below the 75th quantile of 1000 sampled similar dimensional mixing matrices. For practical estimations from finite sample data we use 40 segments, varying the sample size per segment.

Evaluation. ICA methods are often compared in terms of mean correlation coefficient of the estimated sources. Here however, binarization induces heavy noise and individual samples of the estimated sources cannot be accurately estimated. We therefore focus our evaluation on the mixing model, and measure the mean cosine similarity (MCS) of the mixing matrix columns (taking the inherent order and scale indeterminacy of the sources into account, see Appendix D).

6.1 IDENTIFIABILITY

Results. Recall from Section 4 and 5 that the correlations of $\mathbf{q}^u$ convey the information between the binary data and the mixing model, and each of these correlations can be determined from marginal distributions over the corresponding pair of binary observed variables in $\mathbf{x}^u$(in a segment u). Thus, by using the exact pairwise binary distributions of elements of $\mathbf{x}^u$ from Equation 6 as input for BLICA, we are here able to investigate identifiability empirically without any finite sample effects. Figure 3 shows results on which models can be identified when the number of sources and observed variables are equal ($n{=}n_z$). In many cases, the method found the mixing matrix essentially up to machine precision, which can be seen as indication of identifiability. Each box includes 30 different data generating models, for each we ran BLICA 3 times; MCS of the run with highest scaled Gaussian likelihood is plotted. With only 2 segments, or only 2 observed variables (also in Theorem 3), the model is not identifiable in any case. The minimal cases deemed identifiable (up to source scale and order) are ($n{=}5, n_u{=}5$), ($n{=}6, n_u{=}4$), ($n{=}7, n_u{=}4$), ($n{=}8, n_u{=}4$), ($n{=}9, n_u{=}3$), and ($n = 10, n_u = 3$). Thus generally, the more observed variables (n) we have, the less segments (n_u) are needed.

Heuristic Identifiability Analysis. We contrast the results to the well-known heuristic approach to identifiability used in factor analysis. It is based on counting the number of statistics we can calculate (or equations we can form), and the number of unknowns (parameters) we need to solve. If the former is at least as large as the latter, there is hope that the model is identifiable. The calculations in Table 1 are based on Equations 10 and 11 when the number of sources equals the number of observations ($n = n_z$). The statistics correspond to $n_u(n^2 - n)/2$ covariances, $n_u \cdot n$ variances and $n_u \cdot n$ means (for $\mathbf{q}^u$). Unknowns include $n \cdot n$ mixing matrix coefficients, $n_u \cdot n$ (segment-wise) source variances, $n_u \cdot n$ source means, as well as $n_u \cdot n$ scaling terms (diagonal elements of $\mathbf{Q}^u$). In line with the classical literature in factor analysis, we ignore the source order indeterminacy. Figure 3 and Table 1 show a remarkably similar dependence between identifiability and the numbers of the segments and the observed variables: in particular, they agree on the minimal cases identifiable. Interestingly, cases with 2 observed variables as well as the cases with only 2 segments are never identifiable. Note that these computational results together with Section 4 provide a bound for any future analytical results on identifiability. If identifiability turns out to be possible in further cases, e.g., with a different mixing model

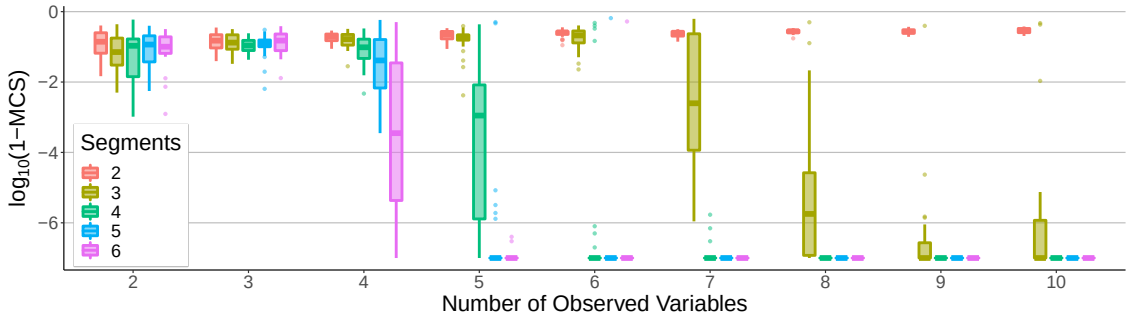

Figure 3: Identifiability with equal number of observed variables and sources. The BLICA method used true (pairwise) probability distributions (i.e. infinite sample limit data). Each box is based on 30 models. Runs with log-error (y-axis) less than -7 (e.g. those that were identified up to machine precision) are marked with -7. Compare to Table 1.

or linking function, the results will need to depend on the particular parametric forms, thus limiting applicability.

Number of Segments	Number of Observed Variables (n)								
	2	3	4	5	6	7	8	9	10
2	-6	-9	-12	-15	-18	-21	-24	-27	-30
3	-7	-9	-10	-10	-9	-7	-4	**0**	**5**
4	-8	-9	-8	-5	**0**	**7**	**16**	27	40
5	-9	-9	-6	**0**	9	21	36	54	75
6	-10	-9	-4	**5**	18	35	56	81	110

Table 1: Heuristic identifiability analysis. Each entry states the number of statistics (equations) minus the number of unknowns. The minimal cases with a non-negative number, suggesting identifiability, are bolded in red.

6.2 FINITE SAMPLE ESTIMATION

Methods. Next we turn our attention to estimation performance from finite sample data. We compare our new BLICA (with different regularization parameter value r) method to its main competitors, fastICA [Himberg and Hyvärinen, 2001, Hyvärinen, 1999] and the baseline implementations of linear iVAE and full MLE. Note that the model of fastICA is somewhat different, but it still employs a linear mixing of the sources and has the same sources scale and order indeterminacies; thus, MCS comparison is sensible. fastICA does not use the segment index, but pools all data from different segments. Recall from Section 5.3 that the linear iVAE uses the same model, but instead of employing the likelihood, it optimizes the ELBO objective through L-BFGS. For runs with $n < 20$ observed variables a time budget of 2h was used, the results that were obtained within the time limit are reported. For larger simulations we allowed 12h per run. To avoid local minima due to the difficult optimization landscape, we ran linear iVAE, full MLE and BLICA with 3 different learning seeds and selected the best run according to the objective function (e.g. likelihood).

Results. Figure 4 (left) shows the result for 10 observed variables and 10 sources. BLICA clearly outperforms others consistently improving with increasing sample size. With smaller dimensions, 6 observed variables and 6 sources in Fig. 4 (center), BLICA needs more samples to achive similar MCS. However, with fewer sources fewer samples are needed: Fig. 4 (right) shows that for 6 observed variables and 2 sources, high MCS can be obtained with only 50 samples per segments. Interestingly, linear iVAE performs well only with fewer sources than observations, while fastICA is not able to reliably estimate the mixing matrix from binary data. Unfortunately, full MLE cannot perform sufficiently many optimization steps within the time limit of 2h even with 6 observed variables in Fig. 4 (center).

Scalability. Figure 5 assesses the performance in higher dimensions over data sets with 40 1000-sample segments, thirty for each n. Only BLICA can estimate the mixing matrix with equal number of observed variables equals and sources in Fig. 5 (left). When the number of sources is fixed to 10 in Fig. 5 (center), also linear iVAE shows improving performance with increasing number of observed variables. Finally, Fig. 5 (right) shows the running time performance of BLICA (Algorithm 1) on the previous runs. The estimation of the quadratic number of correlations starts taking considerable time with 100 observed variables. L-BFGS is relatively quick in solving the optimization problem to a solution close to the final result (i.e. 1% lower MCS), then still gradually improving.

7 RELATED WORK

Our research connects particularly to the following earlier and more recent literature. Himberg and Hyvärinen [2001] consider binary observed vectors $\mathbf{x}$ and binary sources $\mathbf{z}$, so that the ICA mixing model is given by the Boolean expression $x_i = \bigvee_{j=1}^{n_z} a_{ij} \wedge z_j$. They show that this Boolean OR mixing can be approximated by a linear mixing model followed by a unit step function. Thus, they propose to estimate the model by ordinary ICA, and obtain reasonable results when the data is very sparse. Similarly, Nguyen and

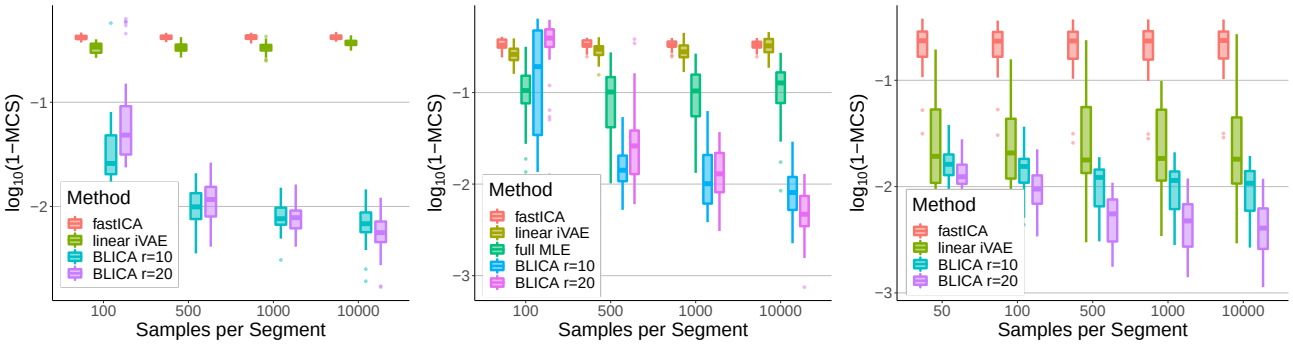

Figure 4: Finite sample performance. Left: 10 observed variables and 10 sources. Center: 6 observed variables and 6 sources. Right: 6 observed variables and 2 sources. Each box is based on thirty 40-segment datasets.

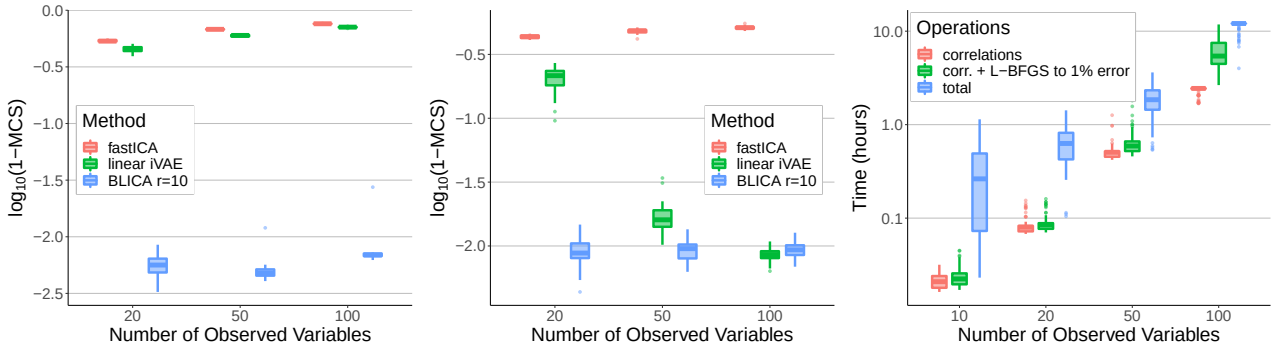

Figure 5: Scalability. Left: equal number of sources and observed variables. Center: 10 sources. Each box is based on thirty 40-segment datasets with 1000 samples per segment. Right: Running times of the steps of BLICA (Algorithm 1).

Zheng [2011] studied binary ICA with OR mixtures by defining a disjunctive generative model. They prove identifiability and propose an algorithm without continuous-valued approximations.

Kabán and Bingham [2006] proposed a model where continuous sources follow a Beta distribution, followed by a binary observation model. While their approach is related to ours, their latent variables are restricted to a finite interval, and they estimate the model using variational approximation which is unlikely to yield consistent estimators. Discrete ICA has further been approached by extensions of LDA where the topic intensities are mutually independent [Podosinnikova et al., 2015, Buntine and Jakulin, 2005, Canny, 2004]. Although their identifiability guarantees are limited [Podosinnikova et al., 2016], their method has the advantage of allowing for discrete data. Lee and Sompolinsky [1999] consider PCA employing a binarized Gaussian model.

Finally, we note that the very idea of estimating latent variable models by non-stationarity, originating in [Matsuoka et al., 1995, Pham and Cardoso, 2001], has been recently increasingly used in estimating generative models [Hyvärinen and Morioka, 2016, Khemakhem et al., 2020] as well as for causal discovery [Zhang et al., 2017, Monti et al., 2019], even in deep learning. Instead of the wide-spread idea of joint diagonalization of covariance matrices [Belouchrani et al., 1997, Tsatsanis and Kweon, 1998], we used correlation matrices without explicit diagonalization criteria; related work on diagonalizing correlation matrices can be found in [Joho and Rahbar, 2002].

8 CONCLUSION

We presented a model for ICA of binary data which is based on a linear latent mixing model and non-stationarity of the sources. We investigated the identifiability, showing some surprising indeterminacies not present in ordinary ICA, including the fact that in the two-variable case the model cannot be identified. We believe that our identifiability results, theoretical and empirical, will be useful in future research on binary ICA. Based on our approach using a Gaussian link function, the likelihood can be obtained in closed form although the Gaussian cumulative distribution function is still computationally heavy. These advances allowed for a practical method BLICA, that combines maximum likelihood estimation and moment-matching; it was shown to be applicable in higher dimensions while still empirically showing consistent behaviour. As future work, we aim to generalize from binary to discrete variables and investigate the potential of the new learning algorithm in applications.

References

Adel Belouchrani, Karim Abed-Meraim, J-F Cardoso, and Eric Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on signal processing*, 45(2):434–444, 1997.

Richard P Brent. *Algorithms for minimization without derivatives*. Courier Corporation, 2013.

Wray Buntine and Aleks Jakulin. Discrete component analysis. In *Proceedings of the 2005 International Conference on Subspace, Latent Structure and Feature Selection*, SLSFS'05. Springer-Verlag, 2005.

John Canny. GaP: A factor model for discrete data. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, page 122–129. Association for Computing Machinery, 2004.

Alan Genz. Comparison of methods for the computation of multivariate normal probabilities. *Computing Science and Statistics*, pages 400–400, 1993.

Alan Genz and Frank Bretz. *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg, 2009.

Johan Himberg and Aapo Hyvärinen. Independent component analysis for binary data: An experimental study. In *Proceedings 3rd International Conference on Independent Component Analysis and Blind Signal Separation, ICA2001*, pages 552–556, 2001.

A. Hyvärinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems (NIPS2016)*, 2016.

Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Networks*, 10(3):626–634, 1999.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*, volume 26. John Wiley & Sons, 2001.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning. In *AISTATS*, volume 89. PMLR, 2019.

Takanori Inazumi, Takashi Washio, Shohei Shimizu, Joe Suzuki, Akihiro Yamamoto, and Yoshinobu Kawahara. Causal discovery in a binary exclusive-or skew acyclic model: Bexsam. *arXiv preprint arXiv:1401.5636*, 2014.

M. Joho and K. Rahbar. Joint diagonalization of correlation matrices by using newton methods with application to blind signal separation. In *Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2002*, pages 403–407, 2002.

Ata Kabán and Ella Bingham. ICA-Based Binary Feature Construction. In *Proceedings of the 6th International Conference on Independent Component Analysis and Blind Signal Separation*, pages 140–148, 2006.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *International Conference on Artificial Intelligence and Statistics, AISTATS*, 2020.

Diederik P. Kingma and Max Welling. Auto-encoding Variational Bayes. In *International Conference on Learning Representations, ICLR*, 2014.

Daniel Lee and Haim Sompolinsky. Learning a continuous hidden variable model for binary data. In *NIPS*, volume 11, 1999.

Xiaopeng Li. Tricks of sigmoid function. `http://eelxpeng.github.io/blog/2017/03/10/Tricks-of-Sigmoid-Function`, 2021. Accessed: 2021-10-5.

Dong C. Liu and Jorge Nocedal. On the Limited Memory BFGS Method for Large Scale Optimization. *Math. Program.*, 45(1-3):503–528, August 1989.

K. Matsuoka, M. Ohya, and M. Kawamoto. A neural net for blind separation of nonstationary signals. *Neural Networks*, 8(3):411–419, 1995.

Jari Miettinen, Klaus Nordhausen, and Sara Taskinen. Blind source separation based on joint diagonalization in r: The packages jade and bssasymp. *Journal of Statistical Software, Articles*, 76(2), 2017.

R. P. Monti, K. Zhang, and A. Hyvärinen. Causal discovery with general non-linear relationships using non-linear ICA. In *Proc. 35th Conf. on Uncertainty in Artificial Intelligence (UAI2019)*, Tel Aviv, Israel, 2019.

Huy Nguyen and Rong Zheng. Binary Independent Component Analysis With or Mixtures. *IEEE Transactions on Signal Processing*, 59(7):3168–3181, 2011. doi: 10.1109/TSP.2011.2144975.

Jonas Peters, Dominik Janzing, and Bernhard Scholkopf. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2436–2450, 2011.

D.-T. Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of nonstationary sources. *IEEE Trans. Signal Processing*, 49(9):1837–1848, 2001.

Anastasia Podosinnikova, Francis Bach, and Simon Lacoste-Julien. Rethinking LDA: Moment matching for discrete ICA. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Anastasia Podosinnikova, Francis Bach, and Simon Lacoste-Julien. Beyond CCA: Moment matching for multi-view models. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 458–467. PMLR, 2016.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *International Conference on Machine Learning, ICML*, 2014.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *J. of Machine Learning Research*, 7:2003–2030, 2006.

Joe Suzuki and Yusuke Inaoka. Causal order identification to address confounding: binary variables. *Behaviormetrika*, pages 1–17, 2021.

M.K. Tsatsanis and Changyeul Kweon. Blind source separation of non-stationary sources using second-order statistics. In *Conference Record of Thirty-Second Asilomar Conference on Signals, Systems and Computers (Cat. No.98CH36284)*, 1998.

Gary R. Waissi and Donald F. Rossin. A sigmoid approximation of the standard normal integral. *Applied Mathematics and Computation*, 77(1):91–95, 1996. ISSN 0096-3003.

David Warton. Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *Journal of the American Statistical Association*, 103:340–349, 02 2008.

Stefan Wilhelm and Manjunath B G. *tmvtnorm: Truncated Multivariate Normal and Student t Distribution*, 2015. R package version 1.4-10.

Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *IJCAI: Proceedings of the Conference*, volume 2017, page 1347. NIH Public Access, 2017.

A PROOF OF THE ROW ORDER INDETERMINACY (THEOREM 1)

Theorem 1. *If the row order of the 2-by-2 mixing matrix* $\mathbf{A}$ *of a binary ICA model is reversed, then the source means* $\boldsymbol{\mu}_{\mathbf{z}}^u$ *and variances* $\boldsymbol{\Sigma}_{\mathbf{z}}^u$ *can be adjusted such that the implied distributions for the observed binary* $\mathbf{x}^u$ *remain identical.*

Proof. Consider two binary ICA models $\mathcal{M} = (\mathbf{A}, \{\boldsymbol{\mu}_{\mathbf{z}}^u\}_u, \{\boldsymbol{\Sigma}_{\mathbf{z}}^u\}_u)$ and $\hat{\mathcal{M}} = (\hat{\mathbf{A}}, \{\hat{\boldsymbol{\mu}}_{\mathbf{z}}^u\}_u, \{\hat{\boldsymbol{\Sigma}}_{\mathbf{z}}^u\}_u)$ that have $n = 2$ observed variables. Let $\hat{\mathbf{A}}$ be $\mathbf{A}$ with rows switched. We define parameters $\{\hat{\boldsymbol{\mu}}_{\mathbf{z}}^u\}_u$, $\{\hat{\boldsymbol{\Sigma}}_{\mathbf{z}}^u\}_u$ and scaling matrices $\{\mathbf{Q}^u\}_u$ such that Equations 10 and 11 in the main paper are satisfied and therefore the binary distributions implied by both models for each segment are identical. First, let $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}}^u = \boldsymbol{\Sigma}_{\mathbf{z}}^u$. This and the row switching of $\mathbf{A}$ means that the covariance matrix of $\mathbf{q}^u$ has just the order switched: $\hat{\boldsymbol{\Sigma}}_{\mathbf{q}}^u[2,2] = \boldsymbol{\Sigma}_{\mathbf{q}}^u[1,1]$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{q}}^u[1,1] = \boldsymbol{\Sigma}_{\mathbf{q}}^u[2,2]$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{q}}^u[1,2] = \boldsymbol{\Sigma}_{\mathbf{q}}^u[1,2]$. The equations implied by Equation 8 in the main paper for each u are:

$$\begin{aligned}
\mathbf{Q}^u[1,1]^2 \boldsymbol{\Sigma}_{\mathbf{q}}^u[1,1] &= \boldsymbol{\Sigma}_{\mathbf{q}}^u[2,2], \\
\mathbf{Q}^u[2,2]^2 \boldsymbol{\Sigma}_{\mathbf{q}}^u[2,2] &= \boldsymbol{\Sigma}_{\mathbf{q}}^u[1,1], \\
\mathbf{Q}^u[1,1] \cdot \mathbf{Q}^u[2,2] \cdot \boldsymbol{\Sigma}_{\mathbf{q}}^u[1,2] &= \boldsymbol{\Sigma}_{\mathbf{q}}^u[1,2].
\end{aligned}$$

These can be solved by setting

$$\begin{aligned}
\mathbf{Q}^u[1,1] &= \sqrt{\boldsymbol{\Sigma}_{\mathbf{q}}^u[2,2]/\boldsymbol{\Sigma}_{\mathbf{q}}^u[1,1]}, \\
\mathbf{Q}^u[2,2] &= \sqrt{\boldsymbol{\Sigma}_{\mathbf{q}}^u[1,1]/\boldsymbol{\Sigma}_{\mathbf{q}}^u[2,2]}.
\end{aligned}$$

Finally, solve for $\hat{\boldsymbol{\mu}}_{\mathbf{q}}^u$ from Equation 10 since $\mathbf{A}, \hat{\mathbf{A}}, \mathbf{Q}^u$ are invertible. □

B PROOF OF THE CORRELATION IDENTIFIABILITY (THEOREM 2)

Theorem 2. *Two binary ICA models imply different distributions for binary observations* $\mathbf{x}^u$ *(in a given segment* u*) if the correlation matrices for* $\mathbf{q}^u$ *are not equal.*

We will first present the result assuming zero means for $\mathbf{q}^u$ since it is more approachable to the reader. Appendix Figure 1 explains this case visually. The full technical proof is given afterwards. Appendix Figures 2 and 3 explain the general case visually.

Proof assuming zero means. We can focus here on bivariate models as the multivariate normal for $\mathbf{q}^u$ can be straightforwardly marginalized to the bivariate case. Suppose the two models respectively imply:

$$\mathbf{q}^u \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{q}}^u), \quad \hat{\mathbf{q}}^u \sim \mathcal{N}(\mathbf{0}, \hat{\boldsymbol{\Sigma}}_{\mathbf{q}}^u), \tag{1}$$

Due to Equations 10 and 11 in the main paper we can also assume we are dealing with "standardized" models where the diagonals of the covariances are units for both models. The correlation/covariance matrices for $\mathbf{q}$ and $\hat{\mathbf{q}}$ are:

$$\boldsymbol{\Sigma}_{\mathbf{q}}^u = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}, \quad \hat{\boldsymbol{\Sigma}}_{\mathbf{q}}^u = \begin{pmatrix} 1 & \beta \\ \beta & 1 \end{pmatrix}.$$

We study the difference in the implied binary distribution by the two models by creating the Gaussian distributions for $\mathbf{q}^u$ and $\hat{\mathbf{q}}^u$ from a single standard multivariate Gaussian source. The distributions can be formed from a standard normal $\mathbf{n} \sim N(\mathbf{0}, \mathbf{I})$, for example by multiplying with matrices

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ \alpha & \sqrt{1-\alpha^2} \end{pmatrix}, \quad \hat{\mathbf{A}} = \begin{pmatrix} 1 & 0 \\ \beta & \sqrt{1-\beta^2} \end{pmatrix}$$

such that

$$\mathbf{q} = \mathbf{A}\mathbf{n}, \quad \hat{\mathbf{q}} = \hat{\mathbf{A}}\mathbf{n}.$$

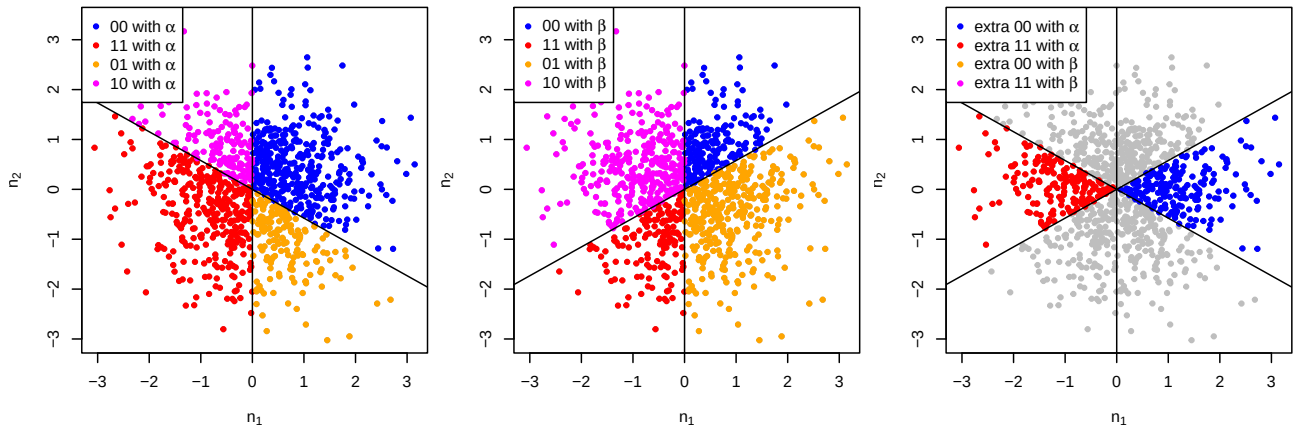

Figure 1: Bivariate standard normal $\mathbf{n}$ and colors indicating which binary assignments are implied with $\alpha = 0.5$ (left) and with $\beta = -0.5$ (center). For this case with zero means, with higher correlation value α we get more 00 and 11 assignments as can be seen from the rightmost plot. Grey points in the rightmost plot do not imply extra 00 or 11 assignments with either correlation value and are irrelevant for the proof.

We will assume $\alpha > \beta$ without loss of generality. Let's look at which values for $\mathbf{n}$ result in different assignments for the binary variables. Recall that the assignment is determined deterministically by the quadrant $\mathbf{q}^u$ and $\hat{\mathbf{q}}^u$ land in. Intuitively, the model with higher correlation α implies more similar values for the binary variables. For the α-model (with $\mathbf{A}$):

$$x_1^u = \begin{cases} 0, & \text{if } n_1 > 0 \\ 1, & \text{if } n_1 < 0 \end{cases}, \quad x_2^u = \begin{cases} 0, & \text{if } -n_2 < \frac{\alpha}{\sqrt{1-\alpha^2}} n_1 \\ 1, & \text{if } -n_2 > \frac{\alpha}{\sqrt{1-\alpha^2}} n_1 \end{cases}.$$

And for the β-model (with $\hat{\mathbf{A}}$):

$$x_1^u = \begin{cases} 0, & \text{if } n_1 > 0 \\ 1, & \text{if } n_1 < 0 \end{cases}, \quad x_2^u = \begin{cases} 0, & \text{if } -n_2 < \frac{\beta}{\sqrt{1-\beta^2}} n_1 \\ 1, & \text{if } -n_2 > \frac{\beta}{\sqrt{1-\beta^2}} n_1 \end{cases}.$$

Note that due to the construction both models agree on the value of the binary variable x_1^u.

With β we get extra assignments such that $x_1^u = x_2^u = 0$ if:

$$n_1 \quad > \quad 0 \quad \text{AND} \quad -n_2 \in [\frac{\alpha}{\sqrt{1-\alpha^2}} n_1, \frac{\beta}{\sqrt{1-\beta^2}} n_1] \tag{2}$$

Since $\alpha > \beta$ and $x/\sqrt{1-x^2}$ is increasing, the interval for n_2 is empty, and no $\mathbf{n}$ implies $x_1^u = x_2^u = 0$ with β if not with α. Suppose $\mathbf{n}$ is such that

$$n_1 \quad > \quad 0 \quad \text{AND} \quad -n_2 \in [\frac{\beta}{\sqrt{1-\beta^2}} n_1, \frac{\alpha}{\sqrt{1-\alpha^2}} n_1].$$

The binary values implied are $x_1^u = x_2^u = 0$ with α and $x_1^u = 0, x_2^u = 1$ with β. Since $\alpha > \beta$ and $x/\sqrt{1-x^2}$ is increasing, the interval for n_2 has non-zero measure. Thus there is a nonzero measure for obtaining extra $x_1^u = x_2^u = 0$ with α. See Figure 1 for pictorial representation of the situation when $\alpha = 0.5$, $\beta = -0.5$. □

Proof. We can focus here on bivariate models as the multivariate normal for $\mathbf{q}^u$ can be straightforwardly marginalized to the bivariate case. Suppose the two models respectively imply:

$$\mathbf{q}^u \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{q}}^u, \boldsymbol{\Sigma}_{\mathbf{q}}^u), \quad \hat{\mathbf{q}}^u \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{\mathbf{q}}^u, \hat{\boldsymbol{\Sigma}}_{\mathbf{q}}^u), \tag{3}$$

Then the marginals are:

$$\begin{aligned} P(x_1^u = 1) &= \Phi(0|\mu_1, \sigma_1^2) = \Phi(-\frac{\mu_1}{\sigma_1}|0,1), \\ P(\hat{x}_1^u = 1) &= \Phi(0|\hat{\mu}_1, \hat{\sigma}_1^2) = \Phi(-\frac{\hat{\mu}_1}{\hat{\sigma}_1}|0,1), \end{aligned}$$

where μ_1, $\hat{\mu}_1$, σ_1, and $\hat{\sigma}_1$ denote the parameters in Equation 3. For the models to imply the same distributions the marginals need to be the same. The same applies for x_2^u with parameters μ_2, $\hat{\mu}_2$, σ_2, and $\hat{\sigma}_2$. Since Φ is monotonically increasing, we can assume from here on:

$$\mu_1\hat{\sigma}_1 = \hat{\mu}_1\sigma_1, \quad \mu_2\hat{\sigma}_2 = \hat{\mu}_2\sigma_2.$$

Due to Equations 10 and 11 in the main paper we can also assume we are dealing with "standardized" models where the diagonals of the covariances are units for both models. We get:

$$\mu_1 = \hat{\mu}_1, \quad \mu_2 = \hat{\mu}_2, \quad \hat{\sigma}_1 = \sigma_1 = \hat{\sigma}_2 = \sigma_2 = 1.$$

The correlation/covariance matrices for $\mathbf{q}$ and $\hat{\mathbf{q}}$ are:

$$\mathbf{\Sigma}_{\mathbf{q}}^u = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}, \quad \hat{\mathbf{\Sigma}}_{\mathbf{q}}^u = \begin{pmatrix} 1 & \beta \\ \beta & 1 \end{pmatrix}$$

We study the difference in the implied binary distribution by the two models by creating the Gaussian distributions for $\mathbf{q}^u$ and $\hat{\mathbf{q}}^u$ from a single standard multivariate Gaussian source. The distributions can be formed from a standard normal $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, for example by multiplying with matrices

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ \alpha & \sqrt{1-\alpha^2} \end{pmatrix}, \quad \hat{\mathbf{A}} = \begin{pmatrix} 1 & 0 \\ \beta & \sqrt{1-\beta^2} \end{pmatrix}$$

such that

$$\mathbf{q} = \mathbf{A}\mathbf{n} + \boldsymbol{\mu}, \quad \hat{\mathbf{q}} = \hat{\mathbf{A}}\mathbf{n} + \hat{\boldsymbol{\mu}},$$

where $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$ due to the earlier. We will assume $\alpha > \beta$ without loss of generality. Let's look at which values for $\mathbf{n}$ result in different assignments for the binary variables. Recall that the assignment is determined deterministically by the quadrant $\mathbf{q}^u$ and $\hat{\mathbf{q}}^u$ land in. Intuitively, the model with higher correlation α implies more similar values for the binary variables. For the α model:

$$x_1^u = \begin{cases} 0, & \text{if } n_1 > -\mu_1 \\ 1, & \text{if } n_1 < -\mu_1 \end{cases}, \quad x_2^u = \begin{cases} 0, & \text{if } -n_2 < \frac{\alpha}{\sqrt{1-\alpha^2}}n_1 + \frac{1}{\sqrt{1-\alpha^2}}\mu_2 \\ 1, & \text{if } -n_2 > \frac{\alpha}{\sqrt{1-\alpha^2}}n_1 + \frac{1}{\sqrt{1-\alpha^2}}\mu_2 \end{cases}.$$

And for the β model:

$$\hat{x}_1^u = \begin{cases} 0, & \text{if } n_1 > -\mu_1 \\ 1, & \text{if } n_1 < -\mu_1 \end{cases}, \quad \hat{x}_2^u = \begin{cases} 0, & \text{if } -n_2 < \frac{\beta}{\sqrt{1-\beta^2}}n_1 + \frac{1}{\sqrt{1-\beta^2}}\mu_2 \\ 1, & \text{if } -n_2 > \frac{\beta}{\sqrt{1-\beta^2}}n_1 + \frac{1}{\sqrt{1-\beta^2}}\mu_2 \end{cases}.$$

Due to the construction both models agree on the value of the binary variable x_1^u.

In the zero mean case presented above we were got more 00 *and* 11 assignments with the higher correlation α than with lower correlation β (Figure 1). Here we can only prove that we always get more 00 *or* 11 assignments, since changing the mean complicated matters (Figure 2 and 3). This is still enough for showing that the distributions are different. First we show that the lower correlation β cannot give extra 00 *and* 11 assignments in comparison to α (separately for positive and negative α).

Case $\alpha > 0$ With β we get additional assignments such that $x_1^u = x_2^u = 0$ if:

$$n_1 > -\mu_1 \quad \text{AND} \quad -n_2 \in [\frac{\alpha}{\sqrt{1-\alpha^2}}n_1 + \frac{1}{\sqrt{1-\alpha^2}}\mu_2, \frac{\beta}{\sqrt{1-\beta^2}}n_1 + \frac{1}{\sqrt{1-\beta^2}}\mu_2] \tag{4}$$

Replacing n_1 with smaller $-\mu_1$ in the lower bound gives a necessary condition for this is:

$$-n_2 \in [-\frac{\alpha}{\sqrt{1-\alpha^2}}\mu_1 + \frac{1}{\sqrt{1-\alpha^2}}\mu_2, \frac{\beta}{\sqrt{1-\beta^2}}n_1 + \frac{1}{\sqrt{1-\beta^2}}\mu_2] \tag{5}$$

With β we get additional assignments $x_1^u = x_2^u = 1$ if:

$$n_1 < -\mu_1 \quad \text{AND} \quad -n_2 \in [\frac{\beta}{\sqrt{1-\beta^2}}n_1 + \frac{1}{\sqrt{1-\beta^2}}\mu_2, \frac{\alpha}{\sqrt{1-\alpha^2}}n_1 + \frac{1}{\sqrt{1-\alpha^2}}\mu_2] \tag{6}$$

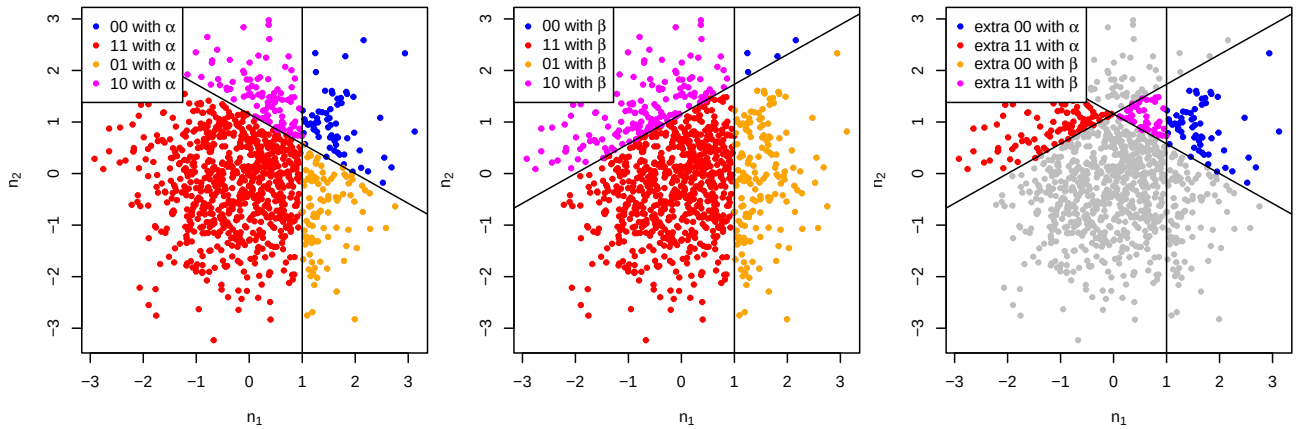

Figure 2: Bivariate standard normal $\mathbf{n}$ and colors indicating which binary assignments are implied with $\alpha = 0.5$ (left) and with $\beta = -0.5$ (center). For this case with $\mu_1 = -1, \mu_2 = -1$, with higher correlation value α we (provably) get more 00 assignments as can be seen from the rightmost plot. Grey points in the rightmost plot do not imply extra 00 or 11 assignments with either correlation value and are irrelevant for the proof.

Replacing n_1 with larger $-\mu_1$ in the upper bound gives a neccessary condition:

$$-n_2 \in [\frac{\beta}{\sqrt{1-\beta^2}}n_1 + \frac{1}{\sqrt{1-\beta^2}}\mu_2, -\frac{\alpha}{\sqrt{1-\alpha^2}}\mu_1 + \frac{1}{\sqrt{1-\alpha^2}}\mu_2] \tag{7}$$

Since the lower bound of Equation 5 matches the upper bound of Equation 7, and the bound is constant with respect to $\mathbf{n}$, both necessary conditions cannot be fulfilled given any fixed model. Therefore the conditions the previous were respectively necessary to, Equation 4 and Equation 6, will not be satisfied either for any fixed model. Note that either Equation 4 or Equation 6 can be satisfied alone.

Case $\alpha < 0$ Also $\beta < 0$ here. With β we get additional assignments such that $x_1^u = x_2^u = 0$ if:

$$n_1 > -\mu_1 \quad \text{AND} \quad -n_2 \in [\frac{\alpha}{\sqrt{1-\alpha^2}}n_1 + \frac{1}{\sqrt{1-\alpha^2}}\mu_2, \frac{\beta}{\sqrt{1-\beta^2}}n_1 + \frac{1}{\sqrt{1-\beta^2}}\mu_2] \tag{8}$$

Replacing βn_1 with larger $-\beta\mu_1$ in the upper bound gives a necessary condition for this is:

$$-n_2 \in [\frac{\alpha}{\sqrt{1-\alpha^2}}n_1 + \frac{1}{\sqrt{1-\alpha^2}}\mu_2, -\frac{\beta}{\sqrt{1-\beta^2}}\mu_1 + \frac{1}{\sqrt{1-\beta^2}}\mu_2] \tag{9}$$

With β we get additional assignments $x_1^u = x_2^u = 1$ if:

$$n_1 < -\mu_1 \quad \text{AND} \quad -n_2 \in [\frac{\beta}{\sqrt{1-\beta^2}}n_1 + \frac{1}{\sqrt{1-\beta^2}}\mu_2, \frac{\alpha}{\sqrt{1-\alpha^2}}n_1 + \frac{1}{\sqrt{1-\alpha^2}}\mu_2] \tag{10}$$

Replacing βn_1 with smaller $-\beta\mu_1$ in the lower bound gives a necessary condition:

$$-n_2 \in [-\frac{\beta}{\sqrt{1-\beta^2}}\mu_1 + \frac{1}{\sqrt{1-\beta^2}}\mu_2, \frac{\alpha}{\sqrt{1-\alpha^2}}n_1 + \frac{1}{\sqrt{1-\alpha^2}}\mu_2] \tag{11}$$

Since the upper bound of Equation 9 matches the lower bound of Equation 11, and the bound is constant with respect to $\mathbf{n}$, both necessary conditions cannot be fulfilled given any fixed model. Therefore the conditions the previous were respectively necessary to, Equation 8 and Equation 10, will not be satisfied either for any fixed model. Note that either Equation 8 or Equation 10 can be satisfied alone.

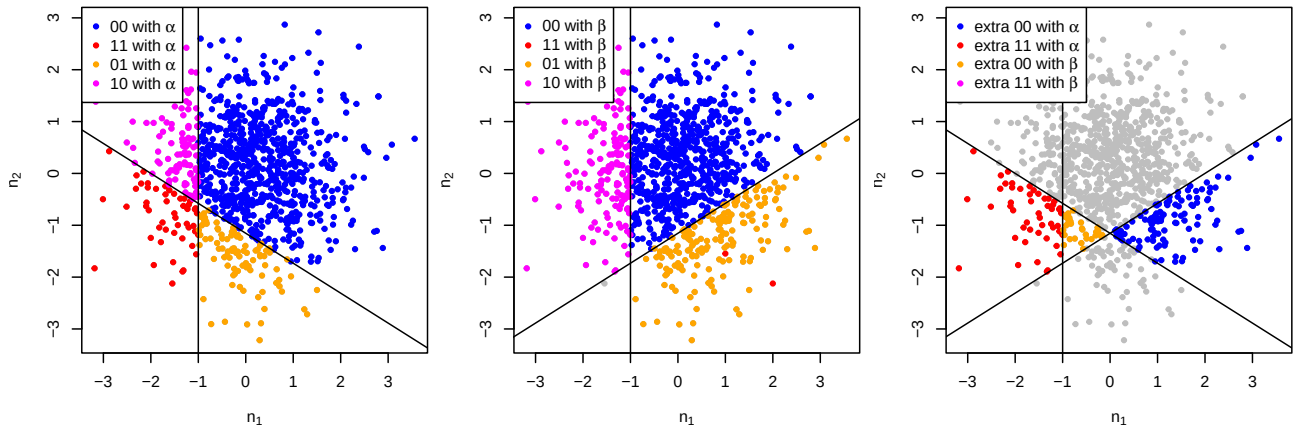

Figure 3: Bivariate standard normal $\mathbf{n}$ and colors indicating which binary assignments are implied with $\alpha = 0.5$ (left) and with $\beta = -0.5$ (center). For this case with $\mu_1 = 1, \mu_2 = 1$, with higher correlation value α we (provably) get more 11 assignments as can be seen from the rightmost plot. Grey points in the rightmost plot do not imply extra 00 or 11 assignments with either correlation value and are irrelevant for the proof.

Extra 00 with α Suppose Equation 4 or Equation 8 is not satisfied. This means that no $\mathbf{n}$ implies $x_1^u = x_2^u = 0$ with β if not with α. Suppose $\mathbf{n}$ is such that

$$\begin{aligned} n_1 &> \max(-\mu_1, \mu_2(\frac{1}{\sqrt{1-\beta^2}} - \frac{1}{\sqrt{1-\alpha^2}})/(\frac{\alpha}{\sqrt{1-\alpha^2}} - \frac{\beta}{\sqrt{1-\beta^2}})) \text{ and} \\ -n_2 &\in [\frac{\beta}{\sqrt{1-\beta^2}} n_1 + \frac{1}{\sqrt{1-\beta^2}}\mu_2, \frac{\alpha}{\sqrt{1-\alpha^2}} n_1 + \frac{1}{\sqrt{1-\alpha^2}}\mu_2]. \end{aligned}$$

The binary values implied are $x_1^u = x_2^u = 0$ with α and $x_1^u = 0, x_2^u = 1$ with β. Furthermore, the following shows that interval for $-n_2$ has non-zero measure. The first multiplication is permitted as the $x/\sqrt{1-x^2}$ is increasing and $\alpha > \beta$.

$$\begin{aligned} n_1 &> \mu_2(\frac{1}{\sqrt{1-\beta^2}} - \frac{1}{\sqrt{1-\alpha^2}})/(\frac{\alpha}{\sqrt{1-\alpha^2}} - \frac{\beta}{\sqrt{1-\beta^2}}) \quad || \cdot (\frac{\alpha}{\sqrt{1-\alpha^2}} - \frac{\beta}{\sqrt{1-\beta^2}}) \\ (\frac{\alpha}{\sqrt{1-\alpha^2}} - \frac{\beta}{\sqrt{1-\beta^2}}) n_1 &> \mu_2(\frac{1}{\sqrt{1-\beta^2}} - \frac{1}{\sqrt{1-\alpha^2}}) \\ \frac{\alpha}{\sqrt{1-\alpha^2}} n_1 &> \frac{\beta}{\sqrt{1-\beta^2}} n_1 + \mu_2(\frac{1}{\sqrt{1-\beta^2}} - \frac{1}{\sqrt{1-\alpha^2}}) \\ \frac{\alpha}{\sqrt{1-\alpha^2}} n_1 + \frac{1}{\sqrt{1-\alpha^2}}\mu_2 &> \frac{\beta}{\sqrt{1-\beta^2}} n_1 + \mu_2(\frac{1}{\sqrt{1-\beta^2}} - \frac{1}{\sqrt{1-\alpha^2}}) + \frac{1}{\sqrt{1-\alpha^2}}\mu_2 \\ &= \frac{\beta}{\sqrt{1-\beta^2}} n_1 + \frac{1}{\sqrt{1-\beta^2}}\mu_2. \end{aligned}$$

Thus there is a nonzero measure for obtaining extra $x_1^u = x_2^u = 0$ with α. See Figure 2 for pictorial representation of the situation when $\alpha = 0.5$, $\beta = -0.5$, $\mu_1 = -1, \mu_2 = -1$.

Extra 11 with α Suppose Equation 6 or Equation 10 is not satisfied. This means that no $\mathbf{n}$ implies $x_1^u = x_2^u = 1$ with β if not with α. Suppose $\mathbf{n}$ is such that

$$\begin{aligned} n_1 &< \min(-\mu_1, \mu_2(\frac{1}{\sqrt{1-\beta^2}} - \frac{1}{\sqrt{1-\alpha^2}})/(\frac{\alpha}{\sqrt{1-\alpha^2}} - \frac{\beta}{\sqrt{1-\beta^2}})) \text{ and} \\ -n_2 &\in [\frac{\alpha}{\sqrt{1-\alpha^2}} n_1 + \frac{1}{\sqrt{1-\alpha^2}}\mu_2, \frac{\beta}{\sqrt{1-\beta^2}} n_1 + \frac{1}{\sqrt{1-\beta^2}}\mu_2]. \end{aligned}$$

The binary values implied are $x_1^u = x_2^u = 1$ with α and $x_1^u = 1, x_2^u = 0$ with β. Furthermore, the following shows that interval for $-n_2$ has non-zero measure. The first multiplication is permitted as the $x/\sqrt{1-x^2}$ is increasing and $\alpha > \beta$.

$$\begin{aligned}
n_1 &< \mu_2\left(\frac{1}{\sqrt{1-\beta^2}} - \frac{1}{\sqrt{1-\alpha^2}}\right)/\left(\frac{\alpha}{\sqrt{1-\alpha^2}} - \frac{\beta}{\sqrt{1-\beta^2}}\right) \quad || \cdot \left(\frac{\alpha}{\sqrt{1-\alpha^2}} - \frac{\beta}{\sqrt{1-\beta^2}}\right) \\
\left(\frac{\alpha}{\sqrt{1-\alpha^2}} - \frac{\beta}{\sqrt{1-\beta^2}}\right)n_1 &< \mu_2\left(\frac{1}{\sqrt{1-\beta^2}} - \frac{1}{\sqrt{1-\alpha^2}}\right) \\
\frac{\alpha}{\sqrt{1-\alpha^2}}n_1 &< \frac{\beta}{\sqrt{1-\beta^2}}n_1 + \mu_2\left(\frac{1}{\sqrt{1-\beta^2}} - \frac{1}{\sqrt{1-\alpha^2}}\right) \\
\frac{\alpha}{\sqrt{1-\alpha^2}}n_1 + \frac{1}{\sqrt{1-\alpha^2}}\mu_2 &< \frac{\beta}{\sqrt{1-\beta^2}}n_1 + \mu_2\left(\frac{1}{\sqrt{1-\beta^2}} - \frac{1}{\sqrt{1-\alpha^2}}\right) + \frac{1}{\sqrt{1-\alpha^2}}\mu_2 \\
&= \frac{\beta}{\sqrt{1-\beta^2}}n_1 + \frac{1}{\sqrt{1-\beta^2}}\mu_2.
\end{aligned}$$

Thus there is a nonzero measure for obtaining extra $x_1^u = x_2^u = 1$ with α. See Figure 3 for pictorial representation of the situation when $\alpha = 0.5$, $\beta = -0.5$, $\mu_1 = 1, \mu_2 = 1$. □

C PROOF OF THEOREM 3

Theorem 3. *If two models $\mathcal{M}$ and $\hat{\mathcal{M}}$ with $n = n_z$ imply the same correlation matrices for $\mathbf{q}^u$ (in a given segment) then the means $\boldsymbol{\mu}_\mathbf{z}^u$ can be adjusted such that the implied binary distributions are identical.*

Proof. If the models imply sample correlations for $\mathbf{q}^u$ they satisfy Equation 11. Thus determine the positive diagonal matrices $\mathbf{Q}^u$ from Equation 11 in the main paper, from the diagonal. Then solve for $\boldsymbol{\mu}_\mathbf{z}^u$ from Equation 10 in the main paper since $\mathbf{A}$ and $\mathbf{Q}^u$ are invertible. Since the equations are satisfied, the implied binary distributions are identical. □

D EVALUATION: MEAN COSINE SIMILARITY

In the binary case, it is more relevant to evaluate the estimated **mixing matrix** than the sources, since the binarization process adds much more noise than simply adding Gaussian noise to the observations. For this purpose, a similar procedure to mean correlation coefficent (MCC) is applied between the estimated mixing matrix and the true mixing matrix.

When there are only two components, the mixing matrix $\mathbf{A} \in \mathbb{R}^{2\times 2}$ can be written considering its column vectors $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2]$. Each vector contains only two elements, so the correlation coefficient cannot be used, since $r(\mathbf{v}_1, \mathbf{v}_2) = 1 \quad \forall\, \mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^2$. In addition, even if $n > 2$, the MCC is undesired because by subtracting the means of each vector, the correlation between "shifted" vectors is the same as if they were not shifted: $r(\mathbf{v}_1 + \mathbf{d}, \mathbf{v}_2) = r(\mathbf{v}_1, \mathbf{v}_2)$ for any $\mathbf{d} \in \mathbb{R}^2$.

Therefore, we employ the **Mean Cosine Similarity (MCS)** instead of the MCC. The MCS uses the cosine similarity – instead of the correlation coefficient – to determine whether the vectors of the true and estimated matrices are aligned:

$$\cos(\mathbf{a}_1, \mathbf{a}_2) = \frac{\mathbf{a}_1 \cdot \mathbf{a}_2}{\|\mathbf{a}_1\|\|\mathbf{a}_2\|} \tag{12}$$

Let us denote the ith column of a matrix $\mathbf{A} \in \mathbb{R}^{n\times n_s}$ as $\mathbf{A}[,i]$. In the MCS calculation, we aim to compare each column of $\mathbf{A}$ with each column of the estimated matrix $\hat{\mathbf{A}}$, thus getting a pair-wise cosine similarity. For simplicity, we consider a column permutation p of matrix $\hat{\mathbf{A}}$ as $\hat{\mathbf{A}}[,p[i]]$. We compute the mean cosine similarity across all the columns for each permutation, and take the maximum, hence defining the MCS as:

$$\text{MCS}(\mathbf{A}, \hat{\mathbf{A}}) = \max_p \left(\frac{1}{n_s} \sum_{i=1}^{n_s} \mid \cos(\mathbf{A}[,i], \hat{\mathbf{A}}[,p[i]]) \mid \right). \tag{13}$$

Instead of actually going through the permutation, the computation can be efficiently performed via a linear assignment problem or a linear program.

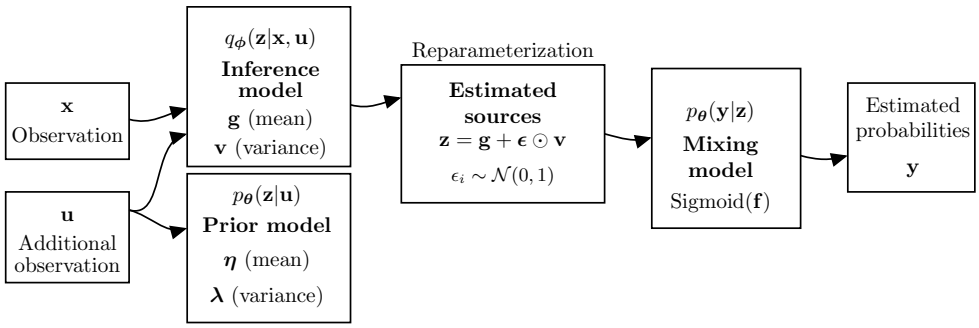

Figure 4: Binary linear iVAE illustration. In VAE terminology: the inference model is equivalent to the encoder, and the mixing model is equivalent to the decoder. The iVAE uses an additionally observed variable $\mathbf{u}$ to estimate the inference model. Additionally, the iVAE estimates a "prior" model for such additionally observed variables. Different from the continuous iVAE, the mixing model does not model the noise explicitly. Also in contrast to the continuous iVAE, the outputs of the model are the estimated probabilities, not the estimated observations. To obtain the probability of each element being 1, a Sigmoid function is applied element-wise to the output of the mixing model. Variables in bold under the model names denote the transformations learned by the model and are described in detail in the text.

E VARIATIONAL AUTOENCODER FOR BINARY DATA (LINEAR IVAE)

Estimation The variational autoencoder[3] iVAE]Khemakhem *et al.*, 2019] aims to estimate the observed data distribution $p(\mathbf{x}|\mathbf{u}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z}|\mathbf{u})d\mathbf{z}$. Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{u}_i)\}_i$, let $q_\mathcal{D}(\mathbf{x}, \mathbf{u})$ be the empirical data distribution. The model learns by maximizing a lower bound $\mathcal{L}$ of the data log-likelihood

$$\mathbb{E}_{q_\mathcal{D}(\mathbf{x},\mathbf{u})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u})] \geq \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}). \tag{14}$$

The loss function is:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) &:= \mathbb{E}_{q_\mathcal{D}(\mathbf{x},\mathbf{u})}[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}|\mathbf{u}) - \log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})]] \\ &= \mathbb{E}_{q_\mathcal{D}(\mathbf{x},\mathbf{u})}[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z},\mathbf{u})] + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})] - \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})]].\end{aligned} \tag{15}$$

To compute the loss function, the expectation over the data distribution is implemented as an average over data samples. In order to deal with expectation over $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$, we use the reparametrization trick and *draw* vectors $\mathbf{z}$ from $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$.

To further develop iVAEs for binary data–which we refer to as linear iVAE in this paper—, we notice that we are working with a factorized Bernoulli observational model. The loss terms developed previously in the continuous iVAE model can remain the same for the inference model and the prior model. However, the loss term referring to the **mixing model** should be modified, since the data follows a **multivariate Bernoulli distribution**. We draw $\mathbf{z}^{(i)} \sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$ using the output of the inference model in the reparameterization trick $\mathbf{z}^{(i)} = \mathbf{g}(\mathbf{x}, \mathbf{u}) + \mathbf{v}(\mathbf{x}, \mathbf{u}) \odot \boldsymbol{\epsilon}^{(i)}$. Thus, the loss term relating to the mixing model can be given as:

$$\begin{aligned}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z},\mathbf{u})] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] &\approx \frac{1}{l}\sum_{j=1}^{l} \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}^{(i)}) = \frac{1}{l}\sum_{i=1}^{l}\sum_{j=1}^{n} \log p_{\boldsymbol{\theta}}(x_j|\mathbf{z}^{(i)}) \\ &= \frac{1}{l}\sum_{i=1}^{l}\sum_{j=1}^{n}\left[x_j \log y_j^{(i)} + (1-x_j)\log(1-y_j^{(i)})\right] \\ &= \frac{1}{l}\sum_{i=1}^{l}\sum_{j=1}^{n} \log \text{Bernoulli}(x_j; y_j^{(i)}),\end{aligned} \tag{16}$$

where y_j is the probability of the observation being 1, $0 \leq y_j \leq 1$, and it is modeled by applying an element-wise sigmoid function to the continuous output of the linear mixing model. Notice that $\mathbf{y}^{(i)}$ is a function of the estimated sources $\mathbf{z}^{(i)}$ drawn from the estimated posterior. Hence, the expectation is approximated by computing the log-probability mass function of a Bernoulli distribution given such probability y_j.

[3]The notation here differs slightly from the previous in order to follow the notation in [Khemakhem *et al.*, 2019] more closely.

Binary model In the model defined, all the transformations are linear, and the sources are drawn from a Gaussian distribution given their segment. Compared to the continuous iVAE, which uses nonlinear transformations in all the models, the binary model is linear and introduces changes to the mixing model and to the prior model. The prior model now estimates not only the log-variances but also the means.

When the observed variables are binary, we use a "Bernoulli MLP" [Kingma and Welling, 2014, Rezende et al., 2014] as a decoder in the mixing model, which aims to estimate parameters from a Bernoulli distribution instead of a Normal distribution. The mixing model is modified from the continuous case by applying a sigmoid function element-wise to the output of the mixing model. In addition, in the binary case, we do not have an explicit factor accounting for the noise in the mixture, as illustrated in Figure 4.

Following, we describe the model in more detail. First of all, we notice that for simplicity and numerical stability when modeling the variances in both the inference model and the prior model, the transformations model the log-variances, which can easily be converted to the variances via exponentiation. With this trick, even a linear transformation can suffice for modeling the log-variances, thus making the model simpler.

The **prior model** is composed of a transformation modeling the prior mean, and a transformation modeling the prior log-variance. The prior **mean** is modeled by

$$\boldsymbol{\eta} : \mathbb{R}^m \to \mathbb{R}^{n_s} \qquad \mathbf{u} \mapsto \boldsymbol{\eta}(\mathbf{u}) \tag{17}$$

where $\boldsymbol{\eta}$ is an affine transformation. So the vector of means is given by $\boldsymbol{\eta}(\mathbf{u}) = \mathbf{W}_\eta \mathbf{u} + \mathbf{b}_\eta$, with matrix weights $\mathbf{W}_\eta \in \mathbb{R}^{n_s \times m}$, and a bias vector $\mathbf{b}_\eta \in \mathbb{R}^{n_s}$. The prior **log-variance** is modeled by

$$\boldsymbol{\lambda} : \mathbb{R}^m \to \mathbb{R}^{n_s} \qquad \mathbf{u} \mapsto \boldsymbol{\lambda}(\mathbf{u}) \tag{18}$$

where $\boldsymbol{\lambda}$ is an affine transformation. The vector of log-variances is given by $\boldsymbol{\lambda}(\mathbf{u}) = \mathbf{W}_\lambda \mathbf{u} + \mathbf{b}_\lambda$, in which $\mathbf{W}_\lambda \in \mathbb{R}^{n_s \times m}$ are the weights, and $\mathbf{b}_\lambda \in \mathbb{R}^{n_s}$ are the biases. Notice that $\boldsymbol{\lambda}$ is unrelated to the notation from the exponential family, since we are modeling both the means and variances.

The **mixing model** learns a transformation

$$\mathbf{f} : \mathbb{R}^{n_s} \to \mathbb{R}^n \qquad \mathbf{z} \mapsto \mathbf{f}(\mathbf{z}) \tag{19}$$

where $\mathbf{f}$ is a linear transformation resulting in the the continuous output $\mathbf{f}(\mathbf{z}) = \mathbf{W}_f \mathbf{z}$, in which $\mathbf{W}_f \in \mathbb{R}^{n \times n_s}$ is the matrix of weights. Then, the probability of the estimated observed variables is given by

$$\mathbf{y} = \text{Sigmoid}(\mathbf{W}_f \mathbf{z}). \tag{20}$$

It is important to notice that each element of $\mathbf{y}$ is an individual probability of the particular observed variable being 1, $\{y_i = P(x_i = 1)\}_{i=1}^n$.

The **inference model** has a transformation modeling the mean, and a transformation modeling the log-variance of the data. The data **mean** is modeled by

$$\mathbf{g} : \mathbb{R}^{n+m} \to \mathbb{R}^{n_s} \qquad (\mathbf{x}, \mathbf{u}) \mapsto \mathbf{g}(\mathbf{x}, \mathbf{u}) \tag{21}$$

where $\mathbf{g}$ is an affine transformation. We denote the concatenation of the vectors $\mathbf{x}$ and $\mathbf{u}$ as $\mathbf{x}||\mathbf{u}$. The vector of means is given by $\mathbf{g}(\mathbf{x}, \mathbf{u}) = \mathbf{W}_g(\mathbf{x}||\mathbf{u}) + \mathbf{b}_g$, for a matrix $\mathbf{W}_g \in \mathbb{R}^{n_s \times (n+m)}$, and a bias vector $\mathbf{b}_g \in \mathbb{R}^{n_s}$. The data **log-variance** is modeled by

$$\mathbf{v} : \mathbb{R}^{n+m} \to \mathbb{R}^{n_s} \qquad (\mathbf{x}, \mathbf{u}) \mapsto \mathbf{v}(\mathbf{x}, \mathbf{u}) \tag{22}$$

where $\mathbf{v}$ is an affine transformation. The vector of log-variances is given by $\mathbf{v}(\mathbf{x}, \mathbf{u}) = \mathbf{W}_v(\mathbf{x}||\mathbf{u}) + \mathbf{b}_v$, where $\mathbf{W}_v \in \mathbb{R}^{n_s \times n+m}$ are the weights and $\mathbf{b}_v \in \mathbb{R}^{n_s}$ the biases.

F FURTHER DETAILS

The experiments were run in computer clusters employing Intel Xeon E5-2680 v4 processors. The running times in Figure 5 (right) in the main paper (as well as all the results in all other experiments) were obtained using a single processor for a specific run.