Binary Independent Component Analysis: A Non-stationarity-based Approach (Supplementary Material)

Antti Hyttinen[1,2] **Vitória Barin-Pacela**[1,2,3] **Aapo Hyvärinen**[1]

[1]Department of Computer Science, University of Helsinki, Helsinki, Finland
[2]Helsinki Institute for Information Technology, Finland
[3]Mila, Université de Montréal, Montréal, Canada

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: $\hat{\boldsymbol{\Sigma}}_{\mathbf{q}}^u[2,2] = \boldsymbol{\Sigma}_{\mathbf{q}}^u[1,1]$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{q}}^u[1,1] = \boldsymbol{\Sigma}_{\mathbf{q}}^u[2,2]$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{q}}^u[1,2] = \boldsymbol{\Sigma}_{\mathbf{q}}^u[1,2]$ (since this matrix is symmetric). The equations implied by Equation 9 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:

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

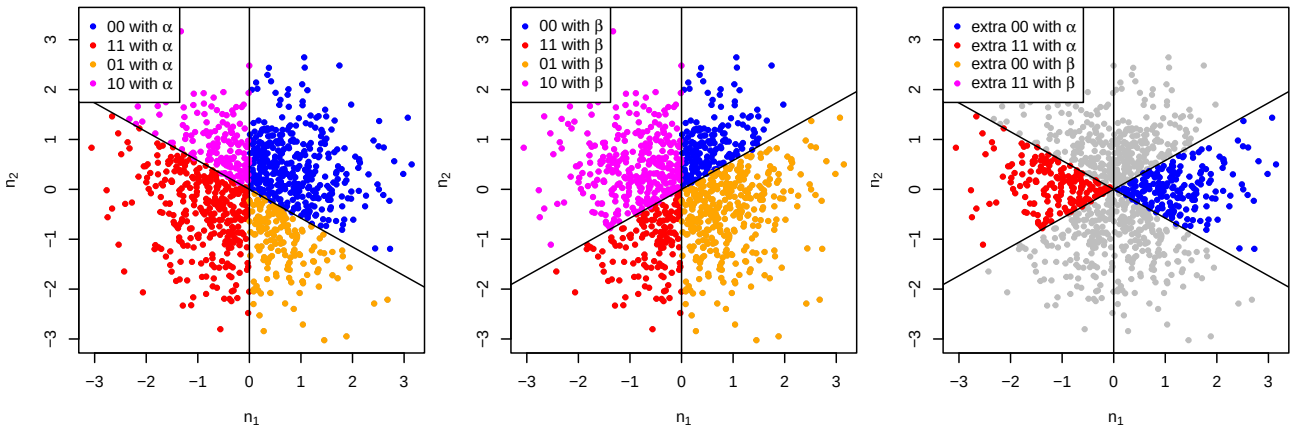

Figure 1: Bivariate standard normal $\mathbf{n}$ and colors indicating which binary assignments are implied with $\alpha = 0.5$ (left) and with $\beta = -0.5$ (center). For this case with zero means, with higher correlation value α we get more 00 and 11 assignments as can be seen from the rightmost plot.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

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:

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

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

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

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 $\mathbf{A}$):

$$x^u_1 = \begin{cases} 0, & \text{if } n_1 > 0 \\ 1, & \text{if } n_1 < 0 \end{cases}, \quad x^u_2 = \begin{cases} 0, & \text{if } -n_2 < \frac{\alpha}{\sqrt{1-\alpha^2}} n_1 \\ 1, & \text{if } -n_2 > \frac{\alpha}{\sqrt{1-\alpha^2}} n_1 \end{cases}.$$

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

$$x^u_1 = \begin{cases} 0, & \text{if } n_1 > 0 \\ 1, & \text{if } n_1 < 0 \end{cases}, \quad x^u_2 = \begin{cases} 0, & \text{if } -n_2 < \frac{\beta}{\sqrt{1-\beta^2}} n_1 \\ 1, & \text{if } -n_2 > \frac{\beta}{\sqrt{1-\beta^2}} n_1 \end{cases}.$$

Note that due to the construction both models agree on the value of the binary variable x^u_1.

With β we get extra assignments such that $x^u_1 = x^u_2 = 0$ if:

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

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

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

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:

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

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:

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

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 $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, 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 got more 00 *and* 11 assignments with the higher correlation α than with the lower correlation β (Figure 1). Here we can only prove that we always get more 00 *or* 11 assignments, since changing the mean complicates matters (Figures 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:

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

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

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

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

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

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

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

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 latter were necessary to, Equation 4 and Equation 6 respectively, 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:

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

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

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

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

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

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

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

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.

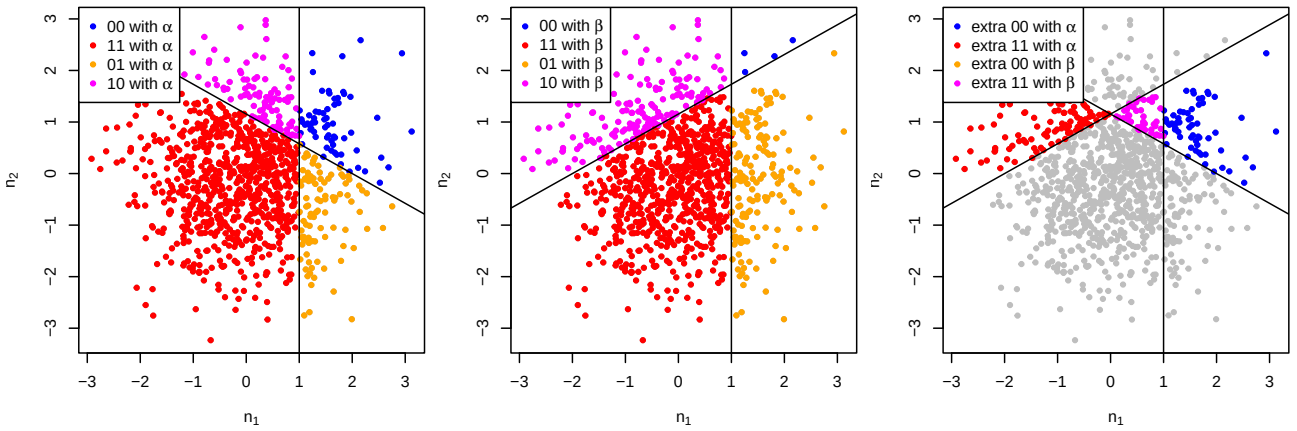

Figure 2: Bivariate standard normal $\mathbf{n}$ and colors indicating which binary assignments are implied with $\alpha = 0.5$ (left) and with $\beta = -0.5$ (center). For this case with $\mu_1 = -1, \mu_2 = -1$, with higher correlation value α we (provably) get more 00 assignments as can be seen from 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.

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

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

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 $\alpha > \beta$.

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

Thus there is a nonzero measure for obtaining extra $x_1^u = x_2^u = 0$ 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

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

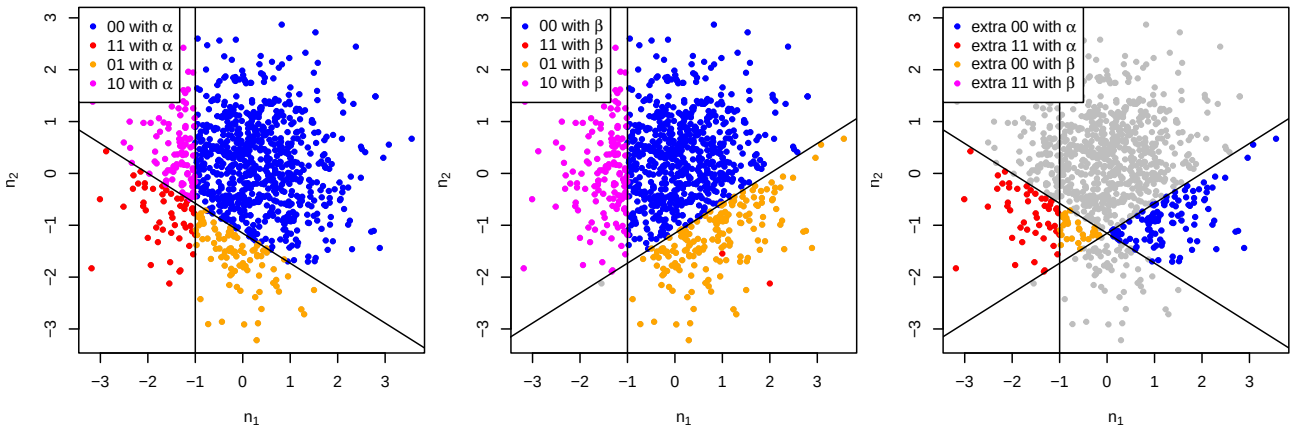

Figure 3: Bivariate standard normal $\mathbf{n}$ and colors indicating which binary assignments are implied with $\alpha = 0.5$ (left) and with $\beta = -0.5$ (center). For this case with $\mu_1 = 1, \mu_2 = 1$, with higher correlation value α we (provably) get more 11 assignments as can be seen from 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.

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 $\alpha > \beta$.

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

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 i^{th} 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.

E VARIATIONAL AUTOENCODER FOR BINARY DATA (LINEAR IVAE)

Estimation The variational autoencoder[1] 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.

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

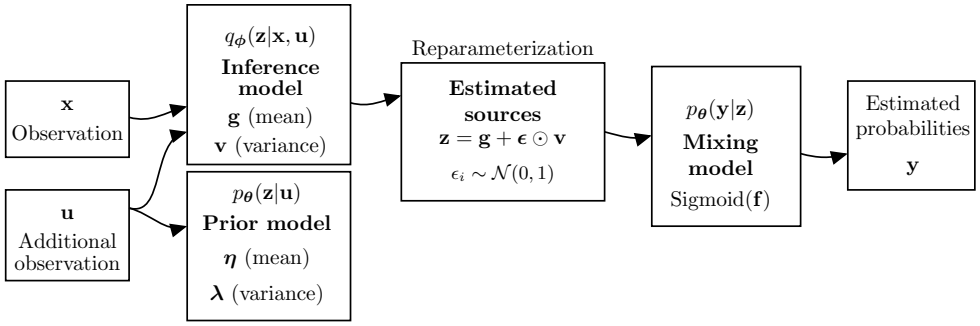

Figure 4: Binary linear iVAE illustration. In VAE terminology: the inference model is equivalent to the encoder, and the mixing model is equivalent to the decoder. The iVAE uses an additionally observed variable $\mathbf{u}$ to estimate the inference model. Additionally, the iVAE estimates a "prior" 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.

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.