

---

# Supplementary Materials and Technical Appendices

---

Yanzhi Chen<sup>1</sup>, Zijing Ou<sup>2</sup>, Adrian Weller<sup>1,3</sup>, Michael Gutmann<sup>4</sup>

<sup>1</sup>University of Cambridge, <sup>2</sup>Imperial College London, <sup>3</sup>Alan Turing Institute, <sup>4</sup>Edinburgh University

## A. Theoretical derivations

### A0. Proof of Theorem 2

**Theorem 2** (MI is vector copula entropy). *The mutual information  $I(X; Y)$  is the negative differential entropy of the vector copula density:*

$$I(X; Y) = -H[c(\mathbf{u}_X, \mathbf{u}_Y)] \quad (1)$$

where  $\mathbf{u}_X$  and  $\mathbf{u}_Y$  are the vector ranks corresponding to  $p(\mathbf{x})$  and  $p(\mathbf{y})$  respectively.

*Proof:* The proof itself relies on the following lemma.

**Lemma 1** (Equivalence between  $p(\mathbf{u}_X, \mathbf{u}_Y)$  and  $c(\mathbf{u}_X, \mathbf{u}_Y)$ ). *The vector copula density  $c(\mathbf{u}_X, \mathbf{u}_Y)$  equals to the probabilistic density function  $p(\mathbf{u}_X, \mathbf{u}_Y)$  of the vector ranks  $\mathbf{u}_X, \mathbf{u}_Y$ .*

*Proof of lemma.* According to the definition of vector ranks, we have the following two identities:

$$p(\mathbf{u}_X) = |J_{\mathbf{x}} \mathbf{u}_X|^{-1} p(\mathbf{x}) = 1, \quad p(\mathbf{u}_Y) = |J_{\mathbf{y}} \mathbf{u}_Y|^{-1} p(\mathbf{y}) = 1$$

where the first equality comes from the law of variable transformation and the second equality comes from the fact that  $p(\mathbf{u}_X) = \mathcal{U}(0, 1)^{d_X}$  and  $p(\mathbf{u}_Y) = \mathcal{U}(0, 1)^{d_Y}$  i.e. they are both factorized uniform distributions. Applying the the law of variable transformation again and rearranging terms, we have

$$p(\mathbf{u}_X, \mathbf{u}_Y) = |J_{\mathbf{x}} \mathbf{u}_X|^{-1} |J_{\mathbf{y}} \mathbf{u}_Y|^{-1} p(\mathbf{x}, \mathbf{y}) = |J_{\mathbf{x}} \mathbf{u}_X|^{-1} |J_{\mathbf{y}} \mathbf{u}_Y|^{-1} p(\mathbf{x}) p(\mathbf{y}) c(\mathbf{u}_X, \mathbf{u}_Y) = c(\mathbf{u}_X, \mathbf{u}_Y)$$

which completes the proof.  $\square$

Now let us turn to the proof of the theorem itself. Due to the bijectivity of vector rank functions (see Definition 1 in the main text), we have

$$I(X; Y) = I(\mathbf{u}_X; \mathbf{u}_Y) = H(\mathbf{u}_X) + H(\mathbf{u}_Y) - H(\mathbf{u}_X, \mathbf{u}_Y) \quad (2)$$

where  $H(\mathbf{u}_X, \mathbf{u}_Y) = H[p(\mathbf{u}_X, \mathbf{u}_Y)]$  is the entropy of the joint distribution  $p(\mathbf{u}_X, \mathbf{u}_Y)$  of the vector ranks  $\mathbf{u}_X, \mathbf{u}_Y$ . The first equality comes from the fact that MI is preserved under diffeomorphic maps  $f, g$  i.e.  $I(X; Y) = I(f(X); g(Y))$ , so that  $I(X; Y) = I(\mathbf{u}_X; \mathbf{u}_Y)$ .

Consider the terms in (2):

- For  $H(\mathbf{u}_X)$  and  $H(\mathbf{u}_Y)$ , we have  $H(\mathbf{u}_X) = H(\mathbf{u}_Y) = 0$  since  $p(\mathbf{u}_X) = \mathcal{U}(0, 1)^{d_X}$  and  $p(\mathbf{u}_Y) = \mathcal{U}(0, 1)^{d_Y}$ ;
- For  $H(\mathbf{u}_X, \mathbf{u}_Y)$ , we have  $H[p(\mathbf{u}_X, \mathbf{u}_Y)] = H[c(\mathbf{u}_X, \mathbf{u}_Y)]$  due to Lemma 1.

Combined, we have  $I(X; Y) = H(\mathbf{u}_X) + H(\mathbf{u}_Y) - H(\mathbf{u}_X, \mathbf{u}_Y) = 0 + 0 - H[c(\mathbf{u}_X, \mathbf{u}_Y)]$ , which completes the proof.  $\square$

### A1. Proof of Proposition 1

**Proposition 1** (Consistency of VCE). *Assuming that (a) the flows  $f_X$  and  $f_Y$  are universal PDF approximator with continuous support and (b) the number of mixture components  $K$  is sufficiently*

large. Define  $\hat{I}_n(X; Y) := \frac{1}{n} \sum_{i=1}^n \log \hat{c}(\hat{\mathbf{u}}_X^{(i)}, \hat{\mathbf{u}}_Y^{(i)})$ . For every  $\epsilon > 0$ , there exists  $n(\epsilon) \in \mathbb{N}$ , such that

$$\left| \hat{I}_n(X; Y) - I(X; Y) \right| < \epsilon, \quad \forall n \geq n(\epsilon), \text{ a.s.}$$

*Proof.* The proof relies on the following lemma.

**Lemma 2** (Consistency of Nested Argmax Estimators). *Let  $\hat{\theta}_1$  be a consistent estimator of  $\theta_1^*$ , and  $\hat{\theta}_2$  is a consistent estimator of  $\operatorname{argmax}_{\theta_2} f(\theta_2, \hat{\theta}_1)$ . Assume that  $f(\theta_1, \theta_2)$  is continuous in both  $\theta_2$  and  $\theta_1$ , and that the maximizer  $\operatorname{argmax}_{\theta_2} f(\theta_2, \theta_1)$  is unique for any  $\theta_1$ . Then  $\hat{\theta}_2$  is also a consistent estimator of  $\operatorname{argmax}_{\theta_2} f(\theta_2, \theta_1^*)$ .*

*Proof of lemma.* Given the consistency of  $\hat{\theta}_1$ , we have  $\hat{\theta}_1 \xrightarrow{P} \theta_1^*$ . By the continuous mapping theorem [1] and the continuity of  $f$ , it follows that

$$f(\theta_2, \hat{\theta}_1) \xrightarrow{P} f(\theta_2, \theta_1^*) \quad \text{for any fixed } \theta_2,$$

which implies that the function  $f(\theta_2, \hat{\theta}_1)$  converges pointwise to  $f(\theta_2, \theta_1^*)$ . Then, by the uniform convergence theorem for maximizers [2], we have

$$\hat{\theta}_2 = \operatorname{argmax}_{\theta_2} f(\theta_2, \hat{\theta}_1) \xrightarrow{P} \operatorname{argmax}_{\theta_2} f(\theta_2, \theta_1^*) = \theta_2^*,$$

which completes the proof.  $\square$

Given the above lemma, we now prove the proposition itself. The complete proof of the proposition consists of four steps:

(a). *Estimation of  $\mathbf{u}_X, \mathbf{u}_Y$  is consistent.* Under the assumption that  $f_x$  and  $f_y$  are universal PDF approximator with continuous supports, they converge to the true marginal distributions in the limit of infinite data. Consequently, the estimated vector ranks  $\hat{\mathbf{u}}_X$  and  $\hat{\mathbf{u}}_Y$  converge in probability to the true vector ranks  $\mathbf{u}_X$  and  $\mathbf{u}_Y$ , respectively. That is,  $\hat{\mathbf{u}}_X \xrightarrow{P} \mathbf{u}_X$  and  $\hat{\mathbf{u}}_Y \xrightarrow{P} \mathbf{u}_Y$ .

(b). *Estimation of  $c$  is consistent given ground truth  $\mathbf{u}_X, \mathbf{u}_Y$ .* By the universal approximation theorem of mixtures [3] and the consistency of maximum likelihood estimator [4], the estimator

$$\operatorname{argmax}_c \frac{1}{m} \sum_{j=1}^m \log c(\mathbf{u}_X, \mathbf{u}_Y), \quad \mathbf{u}_X, \mathbf{u}_Y \sim p(\mathbf{u}_X, \mathbf{u}_Y),$$

is a consistent estimator of the true copula density  $c^*$ . Here  $p(\mathbf{u}_X, \mathbf{u}_Y)$  is the true distribution of vector ranks.

(c). *Estimation of  $c$  is consistent in two-phrase learning.* Combining the results (a)(b), above, by Lemma 2, the estimator

$$\hat{c} = \operatorname{argmax}_c \frac{1}{m} \sum_{j=1}^m \log c(\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y), \quad \hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y \sim \hat{p}(\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y),$$

is also consistent. Here  $\hat{p}$  is the distribution induced by the learned flows.

(d). *Estimation of MI is consistent.* Given the above results, we now show that our estimator is consistent. We begin by defining the following terms:

$$\hat{I}_n(X; Y) := \frac{1}{n} \sum_{i=1}^n \log \hat{c}(\hat{\mathbf{u}}_X^{(i)}, \hat{\mathbf{u}}_Y^{(i)}),$$

$$I'_n(X; Y) := \frac{1}{n} \sum_{i=1}^n \log c^*(\hat{\mathbf{u}}_X^{(i)}, \hat{\mathbf{u}}_Y^{(i)}),$$

$$I_n''(X; Y) := \frac{1}{n} \sum_{i=1}^n \log c^*(\mathbf{u}_X^{(i)}, \mathbf{u}_Y^{(i)}),$$

where  $c^*$  is the true vector copula and  $\mathbf{u}_X, \mathbf{u}_Y$  are the true vector ranks. Note that  $I(X; Y) = \mathbb{E}[\log c^*(\mathbf{u}_X, \mathbf{u}_Y)]$ , which is the limit of  $I_n''(X; Y)$  as  $n \rightarrow \infty$ .

By triangle inequality,

$$\left| I(X; Y) - \hat{I}_n(X; Y) \right| \leq \underbrace{\left| \hat{I}_n(X; Y) - I_n'(X; Y) \right|}_{\Delta} + \underbrace{\left| I_n'(X; Y) - I_n''(X; Y) \right|}_{\nabla} + \left| I_n''(X; Y) - I(X; Y) \right| \quad (3)$$

(i) Since the estimator  $\hat{c}$  is consistent, we know that for every  $\epsilon > 0$ , there exists a sufficiently large  $n \in \mathbb{N}$ , such that  $|\log \hat{c}(\hat{\mathbf{u}}_X^{(i)}, \hat{\mathbf{u}}_Y^{(i)}) - \log c^*(\hat{\mathbf{u}}_X^{(i)}, \hat{\mathbf{u}}_Y^{(i)})| < \epsilon, \forall \hat{\mathbf{u}}_X^{(i)}, \hat{\mathbf{u}}_Y^{(i)}, a.s.$ . Then for the first term in the RHS of (3), we have

$$\Delta = \frac{1}{n} \left| \sum_{i=1}^n \log \hat{c}(\hat{\mathbf{u}}_X^{(i)}, \hat{\mathbf{u}}_Y^{(i)}) - \sum_{i=1}^n \log c^*(\hat{\mathbf{u}}_X^{(i)}, \hat{\mathbf{u}}_Y^{(i)}) \right| \leq \frac{1}{n} \sum_{i=1}^n \left| \log \hat{c}(\hat{\mathbf{u}}_X^{(i)}, \hat{\mathbf{u}}_Y^{(i)}) - \log c^*(\hat{\mathbf{u}}_X^{(i)}, \hat{\mathbf{u}}_Y^{(i)}) \right| \quad (4)$$

$$= \epsilon$$

(ii) Since the estimators  $\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y$  are consistent, we know that for every  $\epsilon > 0$ , there exists a sufficiently large  $n \in \mathbb{N}$ , such that  $|\log c^*(\hat{\mathbf{u}}_X^{(i)}, \hat{\mathbf{u}}_Y^{(i)}) - \log c^*(\mathbf{u}_X^{(i)}, \mathbf{u}_Y^{(i)})| < \epsilon, \forall i, a.s.$ . Then for the second term in the RHS of (3), we have

$$\nabla = \frac{1}{n} \left| \sum_{i=1}^n \log c^*(\hat{\mathbf{u}}_X^{(i)}, \hat{\mathbf{u}}_Y^{(i)}) - \sum_{i=1}^n \log c^*(\mathbf{u}_X^{(i)}, \mathbf{u}_Y^{(i)}) \right| \leq \frac{1}{n} \sum_{i=1}^n \left| \log c^*(\hat{\mathbf{u}}_X^{(i)}, \hat{\mathbf{u}}_Y^{(i)}) - \log c^*(\mathbf{u}_X^{(i)}, \mathbf{u}_Y^{(i)}) \right| \quad (5)$$

$$= \epsilon$$

(iii) For the third term, it vanishes given large  $n$  due to the normal strong law of large numbers under mild conditions.

Given (i)(ii)(iii) and (3), it follows that for every  $\epsilon > 0$ , there exist  $n(\epsilon) \in \mathbb{N}$ , such that  $|\hat{I}_n(X; Y) - I(X; Y)| < \epsilon, \forall n \geq n(\epsilon), a.s.$   $\square$

## A2. Proof of Proposition 2

**Proposition 2** (Error of vector copula-based MI estimate). *Let  $\hat{\mathbf{u}}_X$  and  $\hat{\mathbf{u}}_Y$  be the estimated vector ranks. Let  $p(\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y)$  and  $\hat{p}(\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y)$  be the true and the estimated joint distributions of  $\hat{\mathbf{u}}_X$  and  $\hat{\mathbf{u}}_Y$  respectively<sup>1</sup>. Assuming that sufficient Monte Carlo samples are used to compute  $\hat{I}(X; Y)$  in eq. (5) in the main text, we have*

$$\left| I(X; Y) - \hat{I}(X; Y) \right| \leq \left| H(\hat{\mathbf{u}}_X) + H(\hat{\mathbf{u}}_Y) \right| + KL[p(\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y) \| \hat{p}(\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y)] \quad (6)$$

where the first term on the RHS vanishes as  $\hat{p}(\mathbf{x}) \rightarrow p(\mathbf{x})$  and  $\hat{p}(\mathbf{y}) \rightarrow p(\mathbf{y})$ . In the limit of perfectly learned marginals, we have

$$\left| I(X; Y) - \hat{I}(X; Y) \right| = KL[c \| \hat{c}] \quad (7)$$

where  $c$  and  $\hat{c}$  are the true and estimated vector copula densities respectively.

*Proof.* The proof begins with the following two facts:

- On one hand, due to the bijectivity of flow-based models, we have  $I(X; Y) = I(\hat{\mathbf{u}}_X; \hat{\mathbf{u}}_Y)$ . Then  $I(X; Y) = H(\hat{\mathbf{u}}_X) + H(\hat{\mathbf{u}}_Y) - H(\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y) = H(\hat{\mathbf{u}}_X) + H(\hat{\mathbf{u}}_Y) - \mathbb{E}_p[-\log p(\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y)]$ .

<sup>1</sup>Note that in this case,  $p(\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y)$  is not the true vector copula density unless  $\hat{\mathbf{u}}_X = \mathbf{u}_X$  and  $\hat{\mathbf{u}}_Y = \mathbf{u}_Y$ .

- On the other hand, as  $n \rightarrow \infty$ , we have that by construction,

$$\hat{I}(X; Y) = \mathbb{E}_p[-\log \hat{p}(\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y)].$$

These combined results lead to the following identity:

$$I(X; Y) - \hat{I}(X; Y) = H(\hat{\mathbf{u}}_X) + H(\hat{\mathbf{u}}_Y) - \left( \mathbb{E}_p[-\log p(\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y)] - \mathbb{E}_p[-\log \hat{p}(\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y)] \right) \quad (8)$$

which can be rewritten as

$$I(X; Y) - \hat{I}(X; Y) = H(\hat{\mathbf{u}}_X) + H(\hat{\mathbf{u}}_Y) + KL[p(\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y) \| \hat{p}(\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y)] \quad (9)$$

By applying triangular inequality, we have

$$\left| I(X; Y) - \hat{I}(X; Y) \right| \leq \left| H(\hat{\mathbf{u}}_X) + H(\hat{\mathbf{u}}_Y) \right| + KL[p(\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y) \| \hat{p}(\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y)] \quad (10)$$

which completes the first part of the proof.

Now we turn to the second part of the proof. In the limit of perfectly learned marginals, we have  $\hat{p}(\mathbf{x}) = p(\mathbf{x})$  and  $\hat{p}(\mathbf{y}) = p(\mathbf{y})$ . This yields

$$\hat{\mathbf{u}}_X = \hat{P}(\mathbf{x}) = P(\mathbf{x}) = \mathbf{u}_X, \quad \hat{\mathbf{u}}_Y = \hat{P}(\mathbf{y}) = P(\mathbf{y}) = \mathbf{u}_Y$$

Since  $\mathbf{u}_X \sim \mathcal{U}[0, 1]^{d_X}$  and  $\mathbf{u}_Y \sim \mathcal{U}[0, 1]^{d_Y}$ , we have

$$H(\mathbf{u}_X) = H(\mathbf{u}_Y) = 0.$$

Therefore the first term on the RHS in (10) vanishes.

For the second term on the RHS in (10), since  $\hat{\mathbf{u}}_X = \mathbf{u}_X$  and  $\hat{\mathbf{u}}_Y = \mathbf{u}_Y$ , we have

$$KL[p(\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y) \| \hat{p}(\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y)] = KL[p(\mathbf{u}_X, \mathbf{u}_Y) \| \hat{p}(\mathbf{u}_X, \mathbf{u}_Y)] = KL[c(\mathbf{u}_X, \mathbf{u}_Y) \| \hat{c}(\mathbf{u}_X, \mathbf{u}_Y)]$$

where the last equality comes from Lemma 1, which states that  $p(\mathbf{u}_X, \mathbf{u}_Y) = c(\mathbf{u}_X, \mathbf{u}_Y)$ .

Substituting both terms to (10), we have  $\left| I(X; Y) - \hat{I}(X; Y) \right| = 0 + 0 - KL[c \| \hat{c}] = KL[c \| \hat{c}]. \quad \square$

### A3. Proof of Proposition 3

**Proposition 3** (Vector Gaussian copula as second-order approximation). *A vector Gaussian copula  $c^N$  corresponds to the second-order Taylor expansion of the true vector copula  $c^*$  up to variable transformation.*

*Proof.* Denote  $\mathbf{u} = [\mathbf{u}_X, \mathbf{u}_Y]$  and  $\mathbf{z} = \phi^{-1}(\mathbf{u})$  where  $\phi(\cdot)$  is the element-wise CDF of Gaussian distribution. Let  $p(\mathbf{z})$  be the distribution of  $\mathbf{z}$  and let  $\mu$  be the mode of this distribution. We have

$$\log c^*(\mathbf{u}) = \log |J_{\mathbf{z}} \mathbf{u}|^{-1} + \log p(\mathbf{z}) \quad (11)$$

Applying a second-order Taylor expansion of  $\log p(\mathbf{z})$  around the mode  $\mu$ , we get

$$\log c^*(\mathbf{u}) \approx \log |J_{\mathbf{z}} \mathbf{u}|^{-1} + \log p(\mu) + \mathbf{g}^\top (\mathbf{z} - \mu) + \frac{1}{2} (\mathbf{z} - \mu)^\top \mathbf{H} (\mathbf{z} - \mu)$$

where  $\mathbf{g}$  and  $\mathbf{H}$  is the gradient and the Hessian of  $p(\mathbf{z})$  at  $\mu$ . Since  $\mu$  is the mode, we have  $\mathbf{g} = \mathbf{0}$ . Therefore

$$\log c^*(\mathbf{u}) \approx \log |J_{\mathbf{z}} \mathbf{u}|^{-1} + \underbrace{\log p(\mu) + \frac{1}{2} (\mathbf{z} - \mu)^\top \mathbf{H} (\mathbf{z} - \mu)}_{h(\mathbf{z})}$$

Now consider normalizing this unnormalized (log) density by defining a proper density  $q(\mathbf{z}) = h(\mathbf{z}) / \int h(\mathbf{z}) d\mathbf{z}$ . Given the quadratic form of  $h(\mathbf{z})$ , its corresponding normalized density  $q(\mathbf{z})$  must be a Gaussian distribution with certain mean  $\mu$  and covariance  $\Sigma$ . Then

$$\log c^*(\mathbf{u}) \approx \log |J_{\mathbf{z}} \mathbf{u}|^{-1} + \log \mathcal{N}(\mathbf{z}; \mu, \Sigma) = \log c^N(\mathbf{u}; \mu, \Sigma) \quad (12)$$

Note that RHS itself is a valid probabilistic density function. This shows that the vector Gaussian copula corresponds to the second-order Taylor approximation of the true vector copula in a transformed space induced by CDF of (univariate) standard normal distribution:  $\phi: \mathbb{R} \rightarrow (0, 1)$ .  $\square$

#### A4. Proof of Proposition 4

**Proposition 4** (Vector copula of the product of marginals). *The copula of the distribution  $p'(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$  is a vector Gaussian copula if  $p'(\mathbf{x}, \mathbf{y})$  is absolutely continuous.*

*Proof.* The proof of the proposition relies on the following lemma.

**Lemma 3** (Equivalent representation of vector Gaussian copula). *Let  $f, g$  be two bijective functions. Consider the following data generation process for random variables  $X \in \mathbb{R}^{d_X}$  and  $Y \in \mathbb{R}^{d_Y}$ :*

$$\mathbf{x} = f(\epsilon_{\leq d_X}), \quad \mathbf{y} = g(\epsilon_{> d_X}),$$

$$\epsilon \sim \mathcal{N}(\epsilon; 0, \Sigma)$$

where  $\epsilon \in \mathbb{R}^{d_X + d_Y}$ .  $\epsilon_{\leq d_X}$  denotes the first  $d_X$  dimensions of  $\epsilon$  and  $\epsilon_{> d_X}$  denotes the last  $d_Y$  dimensions of  $\epsilon$ , and  $\mathcal{N}(\epsilon; 0, \Sigma)$  is a Gaussian distribution with zero mean and covariance  $\Sigma$ . Then the vector copula of the distribution  $p(\mathbf{x}, \mathbf{y})$  corresponding to the above generation process is a vector Gaussian copula.

*Proof of lemma:* Let  $f', g'$  be certain bijective functions. The above data generating process can be equivalently expressed as follows:

$$\mathbf{x} = f'(\epsilon'_{\leq d_X}), \quad \mathbf{y} = g'(\epsilon'_{> d_X}),$$

$$\epsilon' \sim \mathcal{N}(\epsilon'; 0, \Sigma')$$

where  $\Sigma' = \begin{bmatrix} \mathbf{I}_X & \Sigma'_{XY} \\ \Sigma'^T_{XY} & \mathbf{I}_Y \end{bmatrix}$  is a p.s.d matrix whose blocks  $\mathbf{I}_X \in \mathbb{R}^{d_X \times d_X}$  and  $\mathbf{I}_Y \in \mathbb{R}^{d_Y \times d_Y}$  are two identity matrices.

Consider  $\mathbf{u}_X = \phi(\epsilon'_{\leq d_X})$  and  $\mathbf{u}_Y = \phi(\epsilon'_{> d_X})$ , where  $\phi$  is the element-wise cumulative distribution function (CDF) of univariate normal distribution. Since different dimensions  $\mathbf{u}_X$  are independent (as dimensions in  $\epsilon'_{\leq d_X}$  are independent), and that each dimension in  $\mathbf{u}_X \sim \mathcal{U}[0, 1]$ ,  $\mathbf{u}_X \sim \mathcal{U}[0, 1]^{d_X}$  and thereby is the vector rank corresponding to  $p(\mathbf{x})$ . Similarly,  $\mathbf{u}_Y$  is also the vector rank corresponding to  $p(\mathbf{y})$ . In summary,  $\mathbf{u}_X$  and  $\mathbf{u}_Y$  are the vector ranks corresponding to  $p(\mathbf{x})$  and  $p(\mathbf{y})$  respectively.

Now consider the joint CDF  $P(\mathbf{u}_X, \mathbf{u}_Y)$  of the random variables  $\mathbf{u}_X$  and  $\mathbf{u}_Y$ :

$$P(\mathbf{u}_X, \mathbf{u}_Y) = P(\epsilon'_{\leq d_X}, \epsilon'_{> d_X}) = \Phi(\epsilon'_{\leq d_X}, \epsilon'_{> d_X}, \Sigma') = \Phi(\phi^{-1}(\mathbf{u}_X), \phi^{-1}(\mathbf{u}_Y), \Sigma')$$

where  $\Phi$  is the CDF of multivariate normal distribution. Comparing the RHS of the equation and the definition of vector Gaussian copula, one can see that  $P(\mathbf{u}_X, \mathbf{u}_Y)$  satisfies the definition of vector Gaussian copula.  $\square$

Given the above lemma, we now turn to the proof of the proposition itself. Literature [5] shows that for any absolutely continuous distribution  $p(\mathbf{x})$ , there exists a diffeomorphism that turns a Gaussian distribution into  $p(\mathbf{x})$ . Then there exist two diffeomorphisms  $f, g$  such that

$$\mathbf{x} \sim p(\mathbf{x}) \Leftrightarrow \mathbf{x} = f(\epsilon_X), \quad \epsilon_X \sim \mathcal{N}(\epsilon_X; 0, \mathbf{I}), \quad \mathbf{y} \sim p(\mathbf{y}) \Leftrightarrow \mathbf{y} = g(\epsilon_Y), \quad \epsilon_Y \sim \mathcal{N}(\epsilon_Y; 0, \mathbf{I})$$

Since  $\mathbf{x} \perp \mathbf{y}$ , we have that

$$I(X; Y) = 0 \Rightarrow I(\epsilon_X; \epsilon_Y) = 0$$

Therefore  $\epsilon_X \perp \epsilon_Y$ . Then

$$p(\epsilon_X, \epsilon_Y) = p(\epsilon_X)p(\epsilon_Y) = \mathcal{N}(\epsilon_X; 0, \mathbf{I})\mathcal{N}(\epsilon_Y; 0, \mathbf{I}) = \mathcal{N}(\epsilon; 0, \mathbf{I})$$

where  $\epsilon = [\epsilon_X, \epsilon_Y]$  is a random variable whose first  $d_X$  dimensions is  $\epsilon_X$  and the last  $d_Y$  dimensions is  $\epsilon_Y$ .

This implies that data  $\mathbf{x}, \mathbf{y} \sim p(\mathbf{x})p(\mathbf{y})$  can be equivalently expressed by the following data generation process:

$$\mathbf{x} = f(\epsilon_X), \quad \mathbf{y} = g(\epsilon_Y),$$

$$\epsilon \sim \mathcal{N}(\epsilon; 0, \mathbf{I})$$

whose vector copula, according to Lemma 3, is a vector Gaussian copula.  $\square$

### A5. Error bound in MI estimation with lossy compression

**Proposition 5** (Error bound in MI estimation with lossy compression). *Let  $X, Y \in \mathbb{R}^D$  be random variables with a joint distribution  $p(\mathbf{x}, \mathbf{y})$  that is absolutely continuous with respect to the Lebesgue measure. Let  $e : \mathbb{R}^D \rightarrow \mathbb{R}^d$  be an encoder and  $h : \mathbb{R}^d \rightarrow \mathbb{R}^D$  be a decoder, both deterministic mappings. Suppose that the conditional log-densities  $\log p(\mathbf{y} | \mathbf{x})$  and  $\log p(\mathbf{y} | \mathbf{x})$  are differentiable w.r.t  $\mathbf{x}$  and  $\mathbf{y}$  respectively, and their gradient are uniformly bounded:*

$$\|\nabla_{\mathbf{x}} \log p(\mathbf{y} | \mathbf{x})\| \leq L, \quad \text{and} \quad \|\nabla_{\mathbf{y}} \log p(\mathbf{x} | \mathbf{y})\| \leq L \quad \forall \mathbf{x}, \mathbf{y}.$$

*Assume the reconstruction error is uniformly bounded:*

$$\|h(e(\mathbf{x})) - \mathbf{x}\|_2 \leq \xi \quad \text{and} \quad \|h(e(\mathbf{y})) - \mathbf{y}\|_2 \leq \xi \quad \forall \mathbf{x}, \mathbf{y}.$$

*Then, as  $\xi \rightarrow 0$ , the mutual information under compression satisfies:*

$$|I(e(X); e(Y)) - I(X; Y)| = O(L\xi).$$

*Proof.* We begin with the following lemma.

**Lemma 4** (Local KL Stability under Uniformly Bounded Score). *Let  $p(y | z)$  be a conditional probability density defined over  $\mathcal{Y} \times \mathcal{Z} \subseteq \mathbb{R}^m \times \mathbb{R}^D$ , and suppose:*

- *For all  $(y, z) \in \mathcal{Y} \times \mathcal{Z}$ , the mapping  $z \mapsto \log p(y | z)$  is differentiable;*
- *The score function is uniformly bounded such that  $\|\nabla_z \log p(y | z)\| \leq L, \forall y \in \mathcal{Y}, z \in \mathcal{Z}$ .*

*Then for any  $z \in \mathcal{Z}$  and any perturbation vector  $\varepsilon \in \mathbb{R}^d$  with  $\|\varepsilon\| \rightarrow 0$ , the KL divergence between nearby conditionals satisfies:*

$$\text{KL}[p(y | z) \| p(y | z + \varepsilon)] = O(L\|\varepsilon\|).$$

*Proof of lemma.* We begin by the Taylor expansion of  $\log p(y | z + \varepsilon)$  around  $z$ :

$$\log p(y | z + \varepsilon) = \log p(y | z) + \nabla_z \log p(y | z)^\top \varepsilon + \underbrace{r(y, \varepsilon)}_{o(\|\varepsilon\|^2)}$$

where  $r(y, \varepsilon)$  is the remainder. Since  $\|\nabla_z \log p(y | z)\| \leq L$ , we have

$$\left| \log p(y | z + \varepsilon) - \log p(y | z) \right| = L\|\varepsilon\| + o(\|\varepsilon\|)$$

Now consider the KL divergence between the two conditional densities:

$$\text{KL}[p(y | z) \| p(y | z + \varepsilon)] = \mathbb{E}_{p(y|z)} \left[ \log \frac{p(y | z)}{p(y | z + \varepsilon)} \right] \leq \mathbb{E} \left[ \left| \log p(y | z) - \log p(y | z + \varepsilon) \right| \right].$$

Substituting the above Taylor expansion term into the KL divergence, we have

$$\text{KL}[p(y | z) \| p(y | z + \varepsilon)] \leq \mathbb{E} \left[ \left| \log p(y | z) - \log p(y | z + \varepsilon) \right| \right] = L\|\varepsilon\| + o(\|\varepsilon\|) = O(L\|\varepsilon\|)$$

which completes the proof of the lemma.  $\square$

To prove the theorem, we need another lemma.

**Lemma 5** (One-side error bound in MI estimation with lossy compression). *Let  $X, Y \in \mathbb{R}^D$  be random variables with a joint distribution  $p(\mathbf{x}, \mathbf{y})$  that is absolutely continuous with respect to the Lebesgue measure. Let  $e : \mathbb{R}^D \rightarrow \mathbb{R}^d$  be an encoder and  $h : \mathbb{R}^d \rightarrow \mathbb{R}^D$  be a decoder, both deterministic mappings. Supposing that all conditions mentioned in Lemma A3 are met. Assume the reconstruction error is uniformly bounded:*

$$\|h(e(\mathbf{x})) - \mathbf{x}\|_2 \leq \xi, \quad \forall \mathbf{x}$$

*Then, as  $\xi \rightarrow 0$ , the mutual information under compression satisfies:*

$$|I(e(X); Y) - I(X; Y)| = O(L\xi).$$

*Proof of lemma.* Denote  $F := h \circ e$  be the reconstruction map, and define the reconstruction residual  $\varepsilon := F(\mathbf{x}) - \mathbf{x}$ . By assumption,  $\|\varepsilon\| \leq \xi$  for all  $\mathbf{x}$ .

We have

$$\begin{aligned}
I(F(X); Y) - I(X; Y) &= \int p(\mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} d\mathbf{x} d\mathbf{y} - \int p(F(\mathbf{x}), \mathbf{y}) \log \frac{p(\mathbf{y}|F(\mathbf{x}))}{p(\mathbf{y})} dF(\mathbf{x}) d\mathbf{y} \\
&= \int p(\mathbf{x}, \mathbf{y}) \log p(\mathbf{y}|\mathbf{x}) d\mathbf{x} d\mathbf{y} - \int p(F(\mathbf{x}), \mathbf{y}) \log p(\mathbf{y}|F(\mathbf{x})) dF(\mathbf{x}) d\mathbf{y} \\
&= \int p(\mathbf{x}, \mathbf{y}) \log p(\mathbf{y}|\mathbf{x}) d\mathbf{x} d\mathbf{y} - \int p(\mathbf{x}, \mathbf{y}) \log p(\mathbf{y}|F(\mathbf{x})) d\mathbf{x} d\mathbf{y} \\
&= \int p(\mathbf{x}) \text{KL} \left[ p(\mathbf{y}|\mathbf{x}) \| p(\mathbf{y}|F(\mathbf{x})) \right] d\mathbf{x} \\
&\leq \sup_{\mathbf{x}} \text{KL} \left[ p(\mathbf{y}|\mathbf{x}) \| p(\mathbf{y}|F(\mathbf{x})) \right]
\end{aligned}$$

By the KL stability lemma, under the assumption that  $|\nabla_{\mathbf{x}} \log p(\mathbf{y} | \mathbf{x})| \leq L, \forall \mathbf{x}$ , we have

$$\left| I(F(X); Y) - I(X; Y) \right| = O(L\|\varepsilon\|) = O(L\xi)$$

where in the last step we substitute  $\|\varepsilon\| \leq \xi$ . □

Given the above lemmas, we now turn to the proof of the proposition itself.

By data process inequality, we have

$$I(X; Y) \geq I(e(X); Y) \geq I(F(X); Y)$$

Therefore

$$0 \leq I(X; Y) - I(e(X); Y) \leq I(X; Y) - I(F(X); Y)$$

hence

$$\left| I(X; Y) - I(e(X); Y) \right| \leq \left| I(X; Y) - I(F(X); Y) \right| = O(L\xi)$$

A similar argument applies to the deviation  $|I(e(X); e(Y)) - I(e(X); Y)|$ , yielding

$$\left| I(e(X); Y) - I(e(X); e(Y)) \right| = O(L\xi)$$

By triangular inequality, we have

$$\left| I(X; Y) - I(e(X); e(Y)) \right| \leq \left| I(X; Y) - I(e(X); Y) \right| + \left| I(e(X); Y) - I(e(X); e(Y)) \right| = O(L\xi)$$

which completes the proof. □

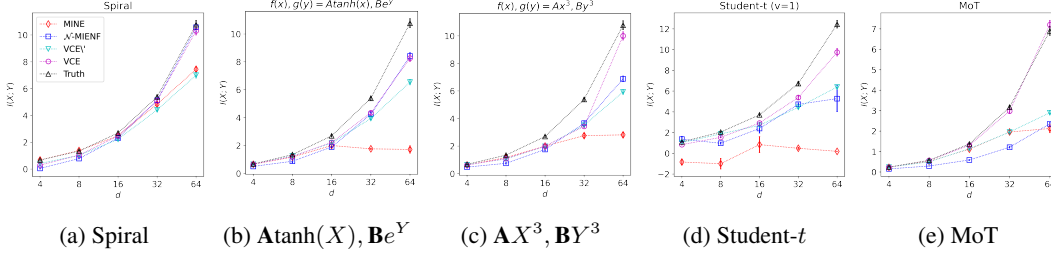


Figure 1: Additional results for the VCE' estimator. In this work, we implement VCE' by taking the reference copula as the independent copula  $c'$ , so that VCE' is equivalent to performing MINE in the vector copula space. Here MoT stands for 'mixture of triangles'.

## B. Experiment details and further results

### B1. Experiment details

**Neural network settings** For controlled experiment, we use the same generative model in  $\mathcal{N}$ -MIENF and our method, and use the same critic network for MINE, InfoNCE and MRE; see below for the details of the networks. All networks are trained by Adam [6] with its default settings, where the learning rate is set to be  $5 \times 10^{-4}$  and the batch size is set to be 512. Early stopping are applied to avoid overfitting in all network training. We use 80% of the data for training and 20% for validation. The detailed architectures of the neural networks used are as follows:

- *Flow models.* We implement the two flow models  $f_X, f_Y$  in our method and  $\mathcal{N}$ -MIENF by a continuous flow model trained by flow matching [7]. This flow model is implemented as a 4-layer MLP with 1024 hidden units per each layer and softplus non-linearity.
- *Critic networks.* We implement the critic network  $f$  in discriminative methods (MINE, MRE and InfoNCE) a MLP with 3 hidden layers, each of which has 500 neurons. A densenet architecture [8] is used for the network, where we concatenate the input of the first layer (i.e.,  $x$  and  $y$ ) and the representation of the penultimate layer before feeding them to the last layer. Leaky ReLU [9] is used as the activation function for all hidden layers.
- *Autoencoders.* For the autoencoder used in part of the experiments, we implement it as a 7-layer MLP with skip connection with architecture  $d_{\text{input}} \rightarrow 512 \rightarrow 512 \rightarrow d_{\text{hidden}} \rightarrow 512 \rightarrow 512 \rightarrow d_{\text{input}}$ , where  $d_{\text{input}}$  and  $d_{\text{hidden}}$  are dimensionalities of the input and the representation respectively.

**Resampling real-world dataset to generate dataset with known MI** We use a technique inspired by that in [10] to turn a real-world dataset  $\mathcal{D}$  with data  $Z \in \mathbb{R}^d$  and ground truth labels  $L \in \{1, 2, \dots, K\}$  into a dataset  $\mathcal{D}'$  with data  $X \in \mathbb{R}^d$  and  $Y \in \mathbb{R}^d$ , where  $I(X; Y)$  is known. The method is based on the assumption  $H[L|Z] \approx 0$  i.e. given the data, there is no ambiguity about its label. This condition is well satisfied for the IMDB dataset [11], where positive and negative comments are well-distinguished [11].

Specifically, to generate data, we first sample  $X, Y, L_X, L_Y \sim p(X|L_X)p(Y|L_Y)p(L_X, L_Y)$  where  $p(L_X, L_Y)$  is a user-defined joint distribution for the discrete random variables  $L_X, L_Y$  and  $p(X|L)$  and  $p(Y|L)$  are the distributions of data within class  $L$ , respectively. It is shown in [10] that under the assumption  $H[L|X] \approx 0$  and  $H[L|Y] \approx 0$ , we have  $I(X; Y) \approx I(L_X; L_Y)$ . The latter is analytically known due to the availability of the discrete distributions  $p(L_X, L_Y)$  and  $p(L_X)p(L_Y)$ .

### B2. Further results and ablation studies

**VCE' performance** In the main text, we introduce an alternative estimator, VCE', which models the copula density  $c$  using a reference copula  $c'$  rather than a mixture of learned vector copulas. In our implementation,  $c'$  is chosen to be the independent copula, and we use the MINE loss to estimate the density ratio  $r = c/c'$ , thereby recovering the target copula as  $c = r \cdot c'$ . As shown in Figure 1, VCE' serves as a useful and reasonable estimator: it significantly outperforms MINE or closely matches its performance across various settings, although it underperforms compared to our main estimator VCE.



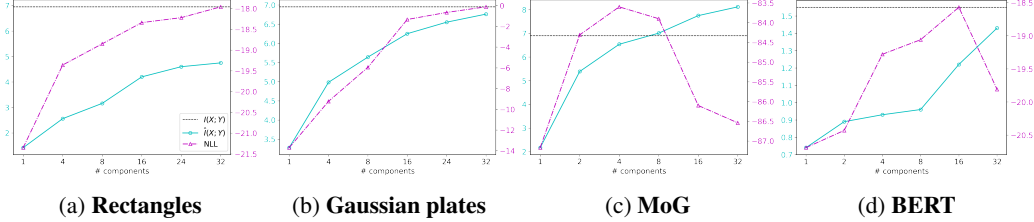


Figure 2: Exploring the effect of number of components  $K$  in the vector copula density  $c$  in the proposed VCE method. The figures shown corresponds to one typical run of the estimator.

	med	std	fail*	$I(X;Y)$		med	std	fail*	$I(X;Y)$
Student- $t$	7.81	5.55	2/10	12.4	Student- $t$	9.50	0.35	0/10	12.4
$\mathbf{A}X^3, \mathbf{B}e^Y$	6.02	0.98	1/10	10.8	$\mathbf{A}X^3, \mathbf{B}e^Y$	8.12	0.12	0/10	10.8

(a) Joint learning

(b) Separate learning

Table 1: Joint learning vs separate learning. Results are collected from 10 independent runs. Data dimensionality is 64. \*Fail: fraction of runs where  $|\hat{I}(X;Y) - I(X;Y)| > \frac{1}{2}I(X;Y)$ . The student- $t$  distribution is with degree of freedom  $\nu = 1$ . The case ' $\mathbf{A}X^3, \mathbf{B}e^Y$ ' corresponds to applying the shown transformation to  $X, Y \sim \mathcal{N}$ , where  $\mathbf{A}$  and  $\mathbf{B}$  are invertible matrices.

**Vector rank computation as data preprocessing** In the main text, we discuss the potential of our vector ranks computation method as a versatile data preprocessing for MI estimation. This is evidenced by the comparison between VCE' and MINE in Figure 1: although both use the same loss function, VCE'—which operates in the vector rank space instead of the original data space—consistently outperforms MINE across various settings. The advantage is especially pronounced in scenarios involving heterogeneous marginals (case.b in Figure 1) and heavy-tailed distributions (case.d in Figure 1). These results demonstrate the effectiveness of vector rank computation as a principled data preprocessing technique for enhancing MI estimation.

**Capacity-complexity trade-off of the copula** A core design in our method is an explicit exploration of the complexity-capacity trade-off of the vector copula. We delve into this process to provide further insights into its impact on the estimation accuracy.

Figure 2 visualizes the model selection procedure described in A. Overall, the negative log-likelihood (NLL) of the vector copula on the validation set generally aligns well with the quality of MI estimate: a higher NLL generally leads to a closer gap between  $I(X;Y)$  and  $\hat{I}(X;Y)$ . Taking the MoG case as example (see Figure 2.c), as the capacity of the vector copula density increases, we observe improvements in both the negative log-likelihood (NLL) and the estimated MI. However, when the copula becomes overly complex, both the NLL and MI estimate worsen. A sweet spot is found at  $K \approx 6$  mixture components in the copula. The results underscore the importance of the complexity-capacity trade-off of the vector copula<sup>2</sup>.

In summary, selecting copula with the best complexity-capacity trade-off is important. The NLL on the validation set serves as an effective criterion in this selection process.

**Joint learning vs separate learning** In addition to the separate *modeling* of marginal distributions and vector copula, an important design of our method is the explicit separation of the *learning* of marginal and copula. We provide empirical evidence to highlight the advantage of this design.

<sup>2</sup>The trends in NLL and MI are not always perfectly aligned. This is reasonable, as the NLL is only calculated on a validation set whereas MI is calculated on the full dataset. This leads to an occasional mismatch between the two values, especially when the validation set is not fully representative of the overall data distribution. Nonetheless, the validation NLL remains a reliable proxy for guiding model selection within our framework.

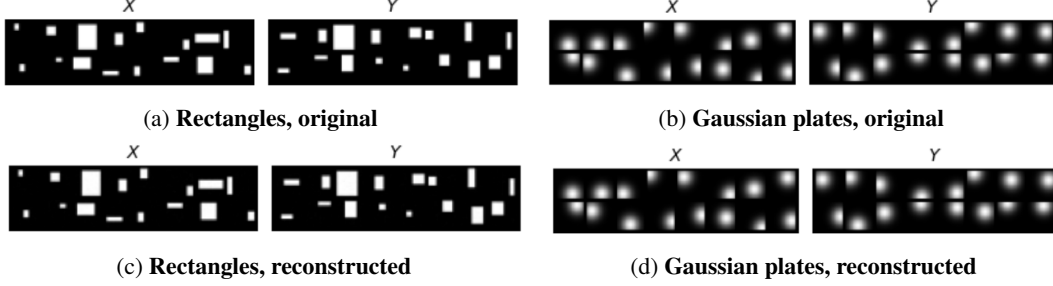


Figure 3: Quality of autoencoder-based compression. Upper panel: original data. Lower panel: reconstructed data with 16-dimensionality latent representation. The compression is near-lossless.

$d_{\text{latent}}$	4	8	16	32	$d_{\text{latent}}$	8	16	32	64
Relative MSE	3e-2	2e-2	8e-3	7e-3	Relative MSE	1e-3	5e-4	3e-4	2e-4

(a) Image - Rectangles

(b) Bert embeddings

Table 2: Quality of autoencoder compression. Relative MSE is defined as  $\mathbb{E}[\|h(e(\mathbf{x})) - \mathbf{x}\|_2^2 / \|\mathbf{x}\|_2^2]$ . Here  $h : \mathbb{R}^{d_{\text{latent}}} \rightarrow \mathbb{R}^{d_{\text{data}}}$  is the decoder. Moderately large  $d_{\text{latent}}$  yields near-lossless compression.

In Table 1, we compare the estimations obtain via joint learning and separate learning on two challenging cases: a 64-dimensional  $t$ -distribution with degree of freedom  $\nu = 1$ , and a distribution with heterogeneous marginal characteristics. As expected, separate learning produces not only more accurate and but also more robust estimation in both cases, as indicated by lower bias and reduced standard deviation. Importantly, we observe that for these two challenging cases, jointly learning the marginal and copula occasionally fails, returning highly biased MI in approximately 2 out of 10 independent runs. This issue does not occur with separate learning. The result highlights the advantage of separate learning in certain cases, which avoids directly learning the marginal distribution and the vector copula altogether — a task that could be otherwise overly challenging.

Beyond accuracy and robustness, separate learning also improves computational efficiency, particularly in the context of model selection. In practice, we observe that separate learning achieves a 2.1~3.7 times acceleration over joint learning. This gain attributes to the fact that we only need to train multiple lightweight models in the copula space, rather than multiple full joint models.

**Quality of autoencoder-based compression** As noted in the main text, we preprocess the image and text datasets using an autoencoder. The quality of this compression is crucial, as highly lossy compression will lead to inaccurate assessment of the performance of different estimators. We investigate the quality of this compression.

Table 2 reports the *relative* mean squared error (Relative MSE) of reconstruction, defined as

$$\mathbb{E}[\|h(e(\mathbf{x})) - \mathbf{x}\|_2^2 / \|\mathbf{x}\|_2^2]$$

where  $e : \mathbb{R}^{d_{\text{data}}} \rightarrow \mathbb{R}^{d_{\text{latent}}}$  is the encoder and  $h : \mathbb{R}^{d_{\text{latent}}} \rightarrow \mathbb{R}^{d_{\text{data}}}$  is the decoder. The results show that reconstruction is nearly perfect for both datasets under the chosen latent dimensionalities ( $d_{\text{latent}} = 16$  for the image dataset and  $d_{\text{latent}} = 32$  for the text dataset), indicating that the compression retains almost all the original information:  $I(X; Y) \approx I(e(X); e(Y))$ , as grounded by Proposition 5 above.

**Comparison to SMILE.** We additionally compare our method to SMILE [12], a robust MI estimator that also provides explicit control over the trade-off between model complexity and capacity, akin to our method. This estimator is defined as

$$\hat{I}(X; Y)_{\text{SMILE}} := \sup_T \mathbb{E}_{p(\mathbf{x}, \mathbf{y})}[T(\mathbf{x}, \mathbf{y})] - \log \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})}[e^{T(\mathbf{x}, \mathbf{y})}],$$

where

$$T(\mathbf{x}, \mathbf{y}) = \text{MLP}(\mathbf{x}, \mathbf{y}).\text{clip}(-\tau, \tau),$$

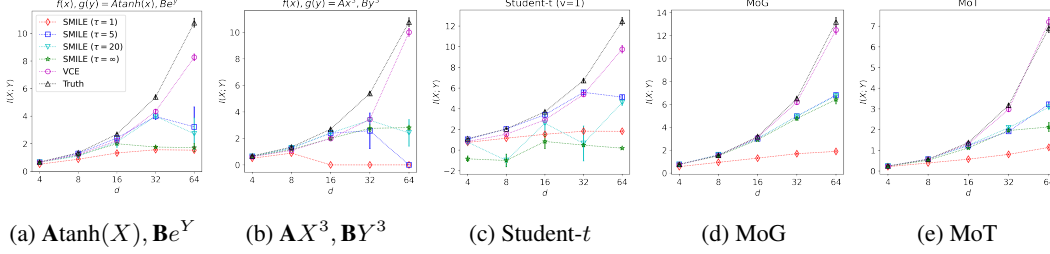


Figure 4: Comparison with the SMILE estimator under different clipping values  $\tau$ .

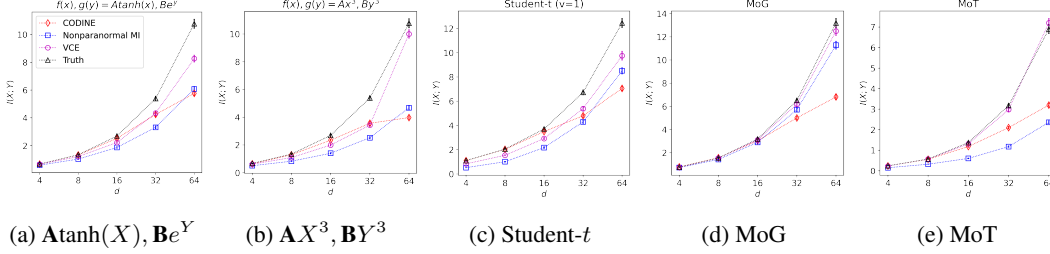


Figure 5: Comparing VCE with classic copula-based estimators e.g. nonparanormal MI (which uses a Gaussian copula to estimate MI) and CODINE (equivalent to classic copula transformation + MINE).

The function  $T : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  is a neural network (typically an MLP) whose output is clipped to the range  $[-\tau, \tau]$ . The clipping parameter  $\tau$  governs the balance between expressiveness and variance:

- A larger  $\tau$  allows the model to capture complex dependencies but increases the estimation variance.
- A smaller  $\tau$  suppresses variance by limiting flexibility, but may reduce the model’s expressiveness.

Figure 4 presents the results, highlighting the superior performance of our proposed VCE method.

**Comparison to classic copula-based MI estimator.** We further compare our method against two *classic* copula-based approaches, which rely on parametric and neural models for copula modeling, respectively:

- Nonparanormal information estimation (*Nonparanormal MI* [13]): This method assumes the data can be approximated by a Gaussian copula model and directly computes MI induced by the corresponding Gaussian copula model.
- Copula neural density estimation (*CODINE* [14]): This method models the copula by a deep neural network and computes MI based on the (classic) copula of the joint distribution and that of the product of marginals.

Figure 5 reports the results. Our proposed VCE estimator consistently outperforms both methods, underscoring the benefits of leveraging vector copulas over classic copula for information estimation.

**Diagnostics on the quality of the estimated vector ranks.** As discussed in the methodology and theory sections, the effectiveness of the proposed VCE method hinges on learning accurate vector ranks. We assess the quality of the estimated ranks  $\hat{\mathbf{u}}_X, \hat{\mathbf{u}}_Y$  from two perspectives:

- *Element-wise uniformity*: Each univariate component  $\hat{\mathbf{u}}_d$  is guaranteed to follow a perfectly uniform distribution in our method, as we employ element-wise empirical ranking when mapping the learned latent in the flow model to  $\mathbf{u}$ .
- *Cross-element independence*: We further examine whether different dimensions,  $\hat{\mathbf{u}}_i$  and  $\hat{\mathbf{u}}_j$ , are statistically independent. Figure 6 visualizes the diagnostic results. In most settings,  $\hat{\mathbf{u}}_i$  and  $\hat{\mathbf{u}}_j$  appear highly independent, with the exception of case (d).

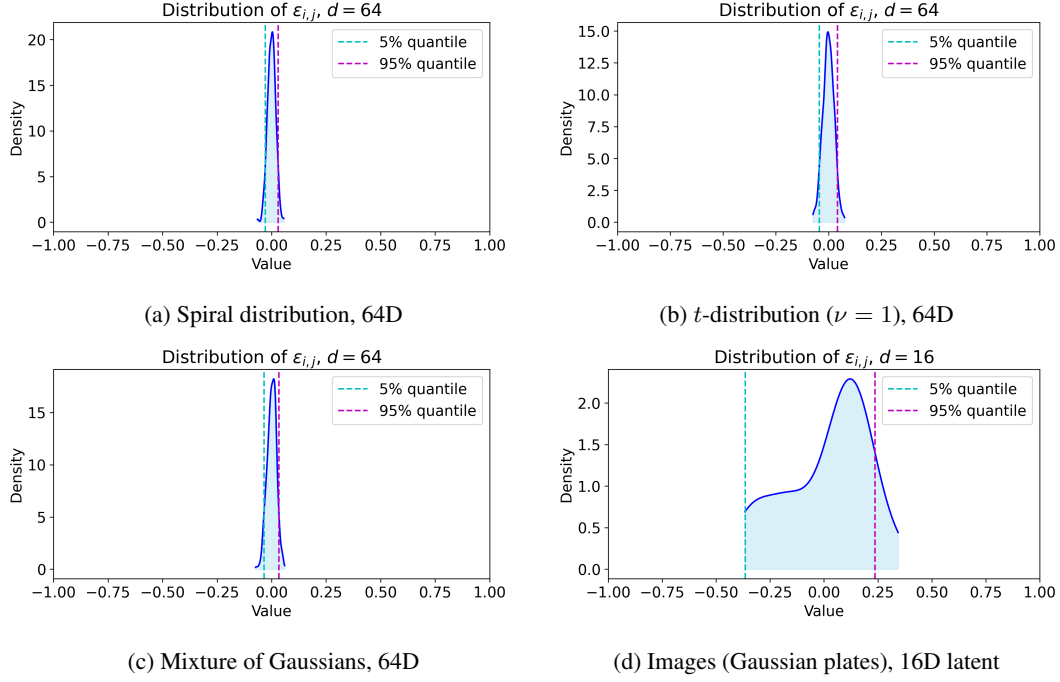


Figure 6: Inspecting the quality of the computed vector ranks. Here, we visualize the distributions of the non-diagonal elements  $\Sigma_{ij}$  in the *correlation matrix*  $\Sigma$  of the estimated vector ranks  $\hat{\mathbf{u}}$ . The results suggest that in most scenarios except case (d),  $\hat{\mathbf{u}}_i$  and  $\hat{\mathbf{u}}_j$  are highly independent.

In practice, in addition to the above visual diagnostic, one can also use the following test statistics  $t(\Sigma)$  to quantify the quality of the learned vector ranks:

$$t(\Sigma) = \max \left( |\mathbb{Q}_{5\%}(\Sigma_{ij})|, |\mathbb{Q}_{95\%}(\Sigma_{ij})| \right)$$

where  $\mathbb{Q}_{\alpha\%}(\Sigma_{ij})$  is the  $\alpha\%$  quantile of the non-diagonal elements in the correlation matrix  $\Sigma$  of  $\hat{\mathbf{u}}$ . Intuitively, a small  $t(\Sigma)$  will indicate that most non-diagonal elements  $\Sigma_{ij}$  in the correlation matrix  $\Sigma$  is close to zero, reflecting strong independence between  $\hat{\mathbf{u}}_i$  and  $\hat{\mathbf{u}}_j$  for different dimensions  $i$  and  $j$ .

Interestingly, we find that even if the vector ranks are not perfectly learned (see e.g. case (d) in Figure 6), our estimator still yields a reasonable estimate. This may be due to that all univariate ranks  $\hat{\mathbf{u}}_d$  are perfectly uniform, so even if  $\hat{\mathbf{u}}_i, \hat{\mathbf{u}}_j$  occasionally exhibit weak dependence, the overall entropy  $|H(\hat{\mathbf{u}})|$  remains low, leading to an acceptable bias in Proposition 5 and a reasonable final estimate.

## References

- [1] Henry B Mann and Abraham Wald. On stochastic limit and order relationships. *The Annals of Mathematical Statistics*, 14(3):217–226, 1943.
- [2] W. Rudin. *Functional Analysis*. International series in pure and applied mathematics. Tata McGraw-Hill, 1974.
- [3] Thomas Hangelbroek and Amos Ron. Nonlinear approximation using gaussian kernels. *Journal of Functional Analysis*, 259(1):203–219, 2010.
- [4] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
- [5] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.
- [6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [7] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2022.
- [8] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [9] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, GA, 2013.
- [10] Gokul Gowri, Xiao-Kang Lun, Allon M Klein, and Peng Yin. Approximating mutual information of high-dimensional variables using learned representations. *arXiv preprint arXiv:2409.02732*, 2024.
- [11] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.
- [12] Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2019.
- [13] Shashank Singh and Barnabás Póczos. Nonparanormal information estimation. In *International Conference on Machine Learning*, pages 3210–3219. PMLR, 2017.
- [14] Nunzio A Letizia, Nicola Novello, and Andrea M Tonello. Copula density neural estimation. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.