InfoNet: Neural Estimation of Mutual Information without Test-Time Optimization

Zhengyang Hu [1] **Song Kang** [2 3] **Qunsong Zeng** [1] **Kaibin Huang** [1] **Yanchao Yang** [1 4]

Abstract

Estimating mutual correlations between random variables or data streams is essential for intelligent behavior and decision-making. As a fundamental quantity for measuring statistical relationships, mutual information has been extensively studied and utilized for its generality and equitability. However, existing methods often lack the *efficiency* needed for real-time applications, such as test-time optimization of a neural network, or the *differentiability* required for end-to-end learning, like histograms. We introduce a neural network called *InfoNet*, which directly outputs mutual information estimations of data streams by leveraging the attention mechanism and the computational efficiency of deep learning infrastructures. By maximizing a dual formulation of mutual information through *large-scale simulated training*, our approach circumvents time-consuming test-time optimization and offers generalization ability. We evaluate the *effectiveness* and *generalization* of our proposed mutual information estimation scheme on various families of distributions and applications. Our results demonstrate that InfoNet and its training process provide a graceful *efficiency-accuracy* trade-off and *order-preserving* properties. Our code and models are available as a comprehensive toolbox to facilitate studies in different fields requiring real-time mutual information estimation.

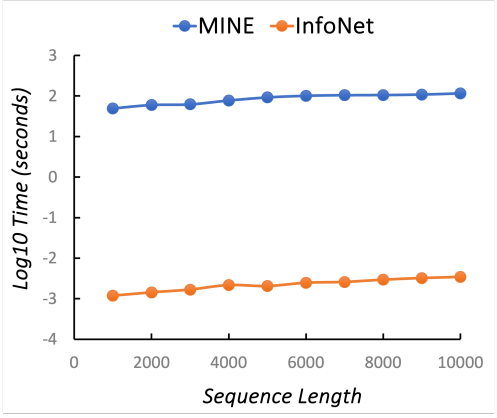


Figure 1: Log-scale run time comparison of MINE (Belghazi et al., 2018a) and the proposed *InfoNet,* which consistently achieves better performance by magnitudes across sequences of varying lengths through bypassing the costly test-time optimization.

1. Introduction

We exist in a universe where various entities are interconnected. At the micro level, particles can exhibit entanglement, as described by quantum mechanics, while at the macro level, celestial bodies are governed by gravity, characterized by general relativity. These interconnections ensure that our observations of the states of different entities around us are intricately correlated rather than independently distributed. This interconnectedness enables us to make informed reasoning and predictions.

Efficiently estimating correlations between scene entities from environmental sensory signals is essential for the emergence of intelligent behavior. This is particularly relevant for embodied agents that interact with the scene and receive large volumes of streaming data, such as video, audio, and touch, within seconds. Rapid correlation estimation helps agents build informative representations of their surroundings and identify crucial elements for survival. Moreover, vast amounts of data are generated every second across the internet, including stock prices, social media messages, e-commerce transactions, and Internet-of-Things devices. Efficiently estimating mutual correlations between differ-

[1]Department of Electrical and Electronic Engineering, the University of Hong Kong [2]School of Information Science and Technology, University of Science and Technology of China [3]Work done as an intern at HKU [4]Musketeers Foundation Institute of Data Science, the University of Hong Kong. Correspondence to: Yanchao Yang <yanchaoy@hku.hk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

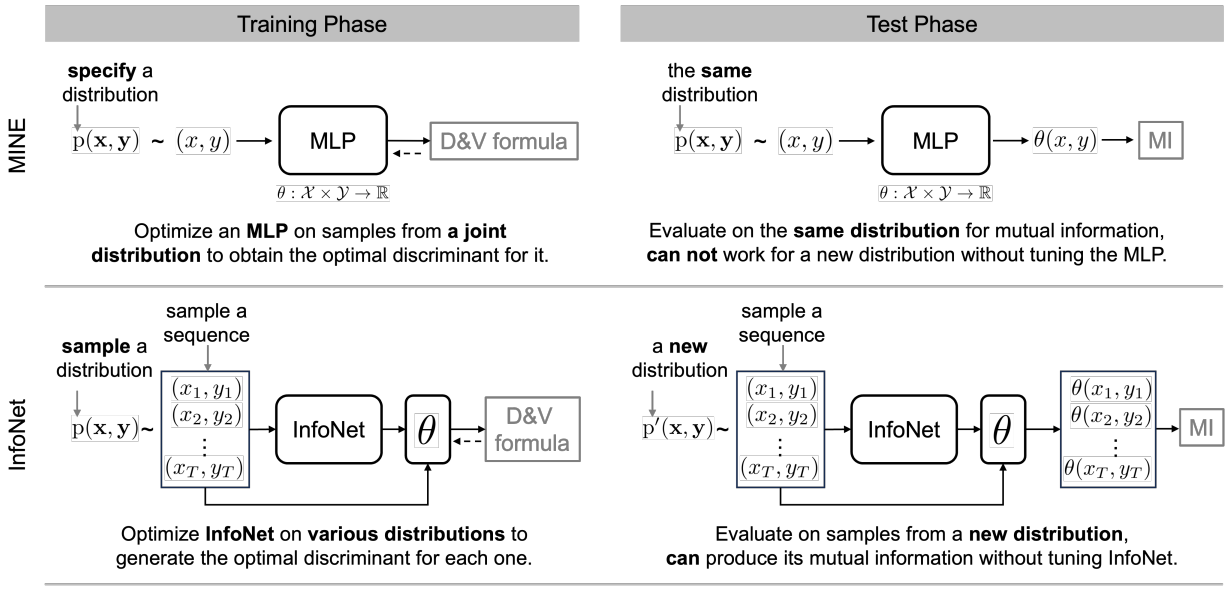


Figure 2: A comparison of MINE (Belghazi et al., 2018a) and the proposed *InfoNet* for neural MI estimation. In the training phase, MINE optimizes an MLP's parameters (as a discriminant function) using the dual formula (Donsker & Varadhan, 1983) against a joint distribution. The optimized MLP then estimates the same distribution's MI with its samples. However, the MLP is not optimal for a new distribution and requires retraining (test-time optimization) before providing an estimate. In contrast, InfoNet is trained to output the optimal discriminant (θ) given samples from various distributions. At test time, InfoNet predicts the optimal discriminant for a new distribution using its samples, leveraging the generalization capability from large-scale training, thus eliminating the need for test-time optimization and increasing efficiency.

ent types or parts of this data informs critical analyses for decision-making.

In this work, we study how to *neuralize* the computation of mutual information (MI) between two random variables from sequences sampled from their *empirical* joint distribution. Specifically, we want to explore whether the estimation of MI can be performed by a neural network without test-time optimization, i.e., taking a pair of sequences as input and speeding out the MI estimate without re-training the network, which guarantees efficiency and differentiability of the estimation procedure.

As a fundamental concept in information theory (Shannon, 1948), a huge amount of effort has been devoted to the estimation of MI (Paninski, 2003; Kraskov et al., 2004), due to its generality and equitability (Reshef et al., 2011; Kinney & Atwal, 2014). For example, many algorithms have been proposed to improve the accuracy and efficiency of MI estimation, which include both non-parametric and parametric methods. However, most of them do not utilize neural networks and can not benefit from advances in deep learning techniques. Recently, MINE (Belghazi et al., 2018a) employs a dual formulation of the Kullback–Leibler divergence and estimates the MI of a pair of sequences by optimizing a neural network's parameters against the dual objective. Even though the estimation can be performed via back-propagation, the optimization process is still behind real-time (Fig. 1, where a joint sequence is sampled from a randomly generated mixture of Gaussian). Moreover, each time the joint distribution changes, a new optimization has to be performed (e.g., the network in MINE is only optimized for a specific distribution, also see Fig. 2, first row), thus not efficient.

To overcome these difficulties, yet still enjoy the *efficiency* of deep networks, we propose a novel network architecture that leverages the attention mechanism (Vaswani et al., 2017) and encodes the aforementioned optimization into the network parameters. Specifically, the proposed network takes as input a sequence of observations (pairs) and outputs a tensor, which aims at maximizing the Donsker-Varadhan (Donsker & Varadhan, 1983) dual and can be converted into an MI estimate by a quick summation over different entries. This way, we *transform* the optimization-based estimation into a feed-forward prediction, thus *bypassing* the time-consuming test-time gradient computation and avoiding sub-optimality via large-scale training on a *wide spectrum* of distributions. Our experiments demonstrate efficiency, accuracy and generalization of the proposed MI neural estimation framework.

In summary, we: 1) propose a neural network and training method for efficiently estimating MI of any distribution

(sequences) without resorting to test-time optimization; 2) conduct an extensive study on the proposed scheme's effectiveness with different distribution families, verifying its accuracy and order-preserving properties; and 3) demonstrate the generalization of the proposed InfoNet on real-world distributions, showcasing promising results in object discovery from videos.

2. Problem Statement

We consider real-world scenarios where an agent receives sensory inputs via multiple channels, i.e., multimodal signals. We treat these observations as random variables and their (synchronized) temporal sequences as if sampled from an empirical joint distribution. More explicitly, we characterize observations $\{(x_t, y_t)\}_{t=1}^{T}$ as samples from a joint distribution $\mathrm{p}(\mathbf{x}, \mathbf{y})$, e.g., by histogramming. Our goal is to compute Shannon's MI between $\mathbf{x}$ and $\mathbf{y}$, i.e., $\mathbb{I}(\mathbf{x}, \mathbf{y})$, in an efficient manner such that an agent can leverage these correlations to learn useful representations and to make effective decisions. Specifically, we aim to train neural networks ϕ such that $\mathcal{C}(\mathbf{x}, \mathbf{y}) = \phi(\{(x_t, y_t)\})$ is an estimation of the MI of $\mathrm{p}(\mathbf{x}, \mathbf{y})$ from the input sequences, without re-training ϕ for different distributions (see Fig. 2, second row). In this work, we focus on the efficient computation of low-dimensional random variables, e.g., 1D/2D, and leverage the projection technique in Goldfeld & Greenewald (2021) for an extension to high-dimensional while maintaining computational efficiency and accuracy.

3. Neural MI Estimation without Test-Time Optimization

MI can be written in Shannon Entropy: $\mathbb{I}(\mathbf{x}, \mathbf{y}) = \mathbb{H}(\mathbf{x}) - \mathbb{H}(\mathbf{x}|\mathbf{y})$, or in Kullback–Leibler divergence: $\mathbb{I}(\mathbf{x}, \mathbf{y}) = D_{\mathrm{KL}}(\mathrm{p}_{\mathbf{x},\mathbf{y}} \| \mathrm{p}_{\mathbf{x}} \cdot \mathrm{p}_{\mathbf{y}})$. However, exact computation is only feasible for discrete variables or a restricted set of distributions (Paninski, 2003). Recently, MINE (Belghazi et al., 2018a) proposes estimating MI using a neural network trained with a dual formula (Donsker & Varadhan, 1983). This method is capable of handling continuous random variables, but requires training from scratch for a different joint distributions $\mathrm{p}'(\mathbf{x}, \mathbf{y})$ (test-time optimization), making real-time MI estimation challenging.

In the following, we provide details of the dual formulation (Donsker & Varadhan, 1983) employed for MI estimation and elaborate on the proposed methods for training the neural network ϕ for computing MI of an unseen distribution without test-time optimization.

Dual Estimation of MI According to Donsker & Varadhan (1983) (also see Gutmann & Hyvärinen (2010)), the KL-divergence between two distributions, p and q, can be written as: $D_{\mathrm{KL}}(\mathrm{p} \| \mathrm{q}) = \sup_{\theta} \mathbb{E}_{\mathrm{p}}[\theta] - \log(\mathbb{E}_{\mathrm{q}}[\exp(\theta)])$, where θ is a discriminant function, whose output is a scalar value, defined on the joint domain with finite expectations. The dual estimation formula for MI is then as follows:

$$\begin{aligned} \mathbb{I}(\mathbf{x}, \mathbf{y}) &= \sup_{\theta} \mathcal{J}^{\mathrm{info}}(\theta; \mathbf{x}, \mathbf{y}) \\ &= \sup_{\theta} \mathbb{E}_{\mathrm{p}_{\mathbf{x},\mathbf{y}}}[\theta] - \log(\mathbb{E}_{\mathrm{p}_{\mathbf{x}} \cdot \mathrm{p}_{\mathbf{y}}}[\exp(\theta)]) \,, \end{aligned} \quad (1)$$

with $\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and $\mathcal{X}, \mathcal{Y}$ the domain of the random variables $\mathbf{x}, \mathbf{y}$ correlated by a joint distribution $\mathrm{p}(\mathbf{x}, \mathbf{y})$. One can instantiate θ as a neural network and train it with the right-hand side in Eq. 1 as the objective, as done in MINE (Belghazi et al., 2018a). The optimal value of the right hand can then serve as the estimate of MI between $\mathbf{x}$ and $\mathbf{y}$ under $\mathrm{p}(\mathbf{x}, \mathbf{y})$. The same training has to be performed for a new distribution, i.e., test-time optimization (see Fig. 2, first row). In contrast, we propose to bypass the test-time optimization by training a novel network architecture that directly outputs the optimal discriminant regarding the dual, using samples from the new distribution. In other words, we treat the *optimal* scalar-valued function θ of a new distribution as the output of the neural network ϕ. This way, we can speed up the estimation by magnitudes and enjoy the benefit of the differentiability of deep neural networks.

Optimal Discriminant Prediction To enable predicting the optimal discriminant θ of a distribution $\mathrm{p}(\mathbf{x}, \mathbf{y})$ from p's samples $\{(x_t, y_t)\}$, we formalize $\theta_{\mathbf{x},\mathbf{y}} = \phi(\{(x_t, y_t)\}) \in \mathbb{R}^{L \times L}$ as a 2D tensor, where L represents the quantization levels of the range of the involved random variables. Now, the value of $\theta_{\mathbf{x},\mathbf{y}}(x_t, y_t)$ for a continuous pair (x_t, y_t) can be directly read out from the tensor as a look-up table with correct indexing and appropriate interpolation.

To facilitate the prediction, we design a neural network by adapting the Perceiver IO from (Jaegle et al., 2021). The proposed network structure ϕ is illustrated in Fig. 3 and named as *InfoNet*. It takes in a pair of jointly sampled sequences, e.g., $\{(x_t, y_t)\}_{t=1}^{T} \in \mathbb{R}^{T \times 2}$, and outputs a tensor $\theta_{\mathbf{x},\mathbf{y}}$ as a discretization of the scalar function θ in Eq. 1.

The input is initially processed through two distinct pathways. The first pathway involves processing joint samples, which are encoded by an attention module and mapped to a tensor in $\mathbb{R}^{M \times D}$, where M, D are the number and dimension of the learnable queries. This tensor is further processed by a series of self-attention modules to produce a set of keys and values. The second pathway involves processing individual variables X and Y. Each sample of X or Y is mapped separately using a multi-layer perceptron (MLP), then consumed by an attention module to produce an output of shape $\mathbb{R}^{L \times L}$, which represents a set of queries. Notably, each query is a combination of the Y samples and can serve as a token to computing the marginals $\theta_{\mathbf{x},\mathbf{y}}(:, y)$. Furthermore, the outputs from the two pathways are then

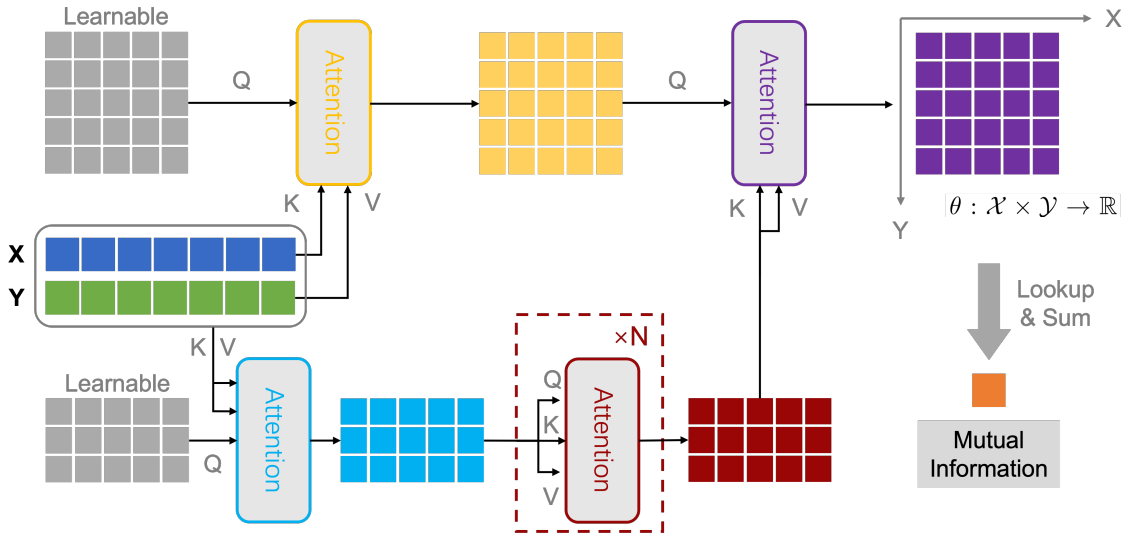


Figure 3: The proposed InfoNet architecture for MI prediction comprises learnable queries and attention blocks. It accepts a sequence of samples from two random variables and outputs a look-up table (top-right) representing a discretization of the optimal scalar discriminant function defined on the joint domain in the Donsker-Varadhan representation (Donsker & Varadhan, 1983). The MI between the two random variables (sequences) can then be calculated by summation according to Eq. 1. Note that the input sequences for training are sampled from various distributions. Please also refer to Fig. 2 for a comparison between MINE and InfoNet training schemes.

integrated by an attention module to generate a 2D tensor $\theta_{\mathbf{x},\mathbf{y}}$ of shape $\mathbb{R}^{L\times L}$, which serves as the look-up table. Finally, a convolutional layer with a non-learnable Gaussian kernel is applied to $\theta_{\mathbf{x},\mathbf{y}}$ to enhance the smoothness.

In addition, to improve the training efficiency, we apply a copula transformation Durante & Sempi (2010) on the sequences before inputting them to the network, which helps normalize their range to $[0, 1]$. It is worth noting that the invariance property of MI under bijective mappings of the RVs ensures that such a transformation does not change the MI between the two sequences. More details on this copula transformation can be found in Appendix A.1.

With the predicted (discretized) discriminant function $\theta_{\mathbf{x},\mathbf{y}}$, we can then compute an estimate of the MI between $\mathbf{x}$ and $\mathbf{y}$ using the quantity $\mathcal{J}^{\text{info}}(\theta; \mathbf{x}, \mathbf{y})$ in Eq. 1.

To ensure that the predicted discriminant is optimal for $\mathrm{p}(\mathbf{x}, \mathbf{y})$ under Eq. 1, we train the neural network ϕ using the following objective (a discretization of Eq. 1):

$$\begin{aligned}\mathcal{L}_{\text{MI}}(\phi, \mathcal{D}) &= \frac{1}{N}\sum_{i=1}^{N} \mathcal{J}(\theta_{\mathbf{x}^i,\mathbf{y}^i}; \mathbf{x}^i, \mathbf{y}^i) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{T}\sum_{t=1}^{T}\theta_{\mathbf{x}^i,\mathbf{y}^i}(x_t^i, y_t^i)\right. \\ &\quad \left. - \log\left(\frac{1}{T}\sum_{t=1}^{T}\exp(\theta_{\mathbf{x}^i,\mathbf{y}^i}(x_t^i, \tilde{y}_t^i))\right)\right\}. \quad (2)\end{aligned}$$

Here $\mathcal{D}$ is a dataset of N *different distributions*, i.e., $\mathcal{D} = \{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^N$ with each $(\mathbf{x}^i, \mathbf{y}^i) = \{(x_t^i, y_t^i)\}_{t=1}^T$ representing a sequence sampled from a distribution $\mathrm{p}^i(\mathbf{x}, \mathbf{y})$. Please also note that the second expectation in Eq. 1 is over the product of the marginals, so we write $\tilde{y}_i$ in the second summation in Eq. 2 to emphasize the difference. The marginal distribution $\mathrm{p}^i(\mathbf{y})$ can be sampled by simply breaking the pairing between x_t^i and y_t^i, e.g., by shuffling $\{y_t^i\}_{t=1}^T$ as if x does not exist. We detail the generation of the training samples of many different distributions in Sec. 4. Since the training of ϕ (InfoNet) is performed with a large set of (simulated) distributions between $\mathbf{x}$ and $\mathbf{y}$ instead of a single distribution as in MINE (Belghazi et al., 2018a), the predicted $\theta_{\mathbf{x}',\mathbf{y}'}$ from samples $\{(x_t', y_t')\}_{t=1}^T$ of a new distribution $\mathrm{p}'(\mathbf{x}, \mathbf{y})$ is supposed to maximize the quantity $\mathcal{J}^{\text{info}}$ in Eq. 1 through generalization. We also verify the generalization of the trained InfoNet with an extensive study in the experimental section.

4. Data Generation and Training Algorithm

To generate training data, we consider sampling the joint distributions (sequences) $\mathcal{D} = \{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^N$ from Gaussian Mixture Models (GMMs). It is widely accepted that GMM is a versatile and effective tool for modeling real-world distributions due to its capability to handle complex and noisy data (Reynolds et al., 2009). Specifically, GMMs represent a family of distributions as a weighted sum of Gaus-

Algorithm 1 InfoNet Training

Require: A maximum number of Gaussian components; Learning rate η

repeat
  Randomly select N two-dimensional Gaussian mixture distributions
  Select T data points from each joint distribution
  Shuffle the $\mathbf{y}$ component to get its marginal samples
  Put joint samples into the model and get N two-dimensional lookup tables $\theta_{\mathbf{x}^i,\mathbf{y}^i}$'s
  Apply lookup function to get the corresponding discriminant values $\theta_{\mathbf{x}^i,\mathbf{y}^i}(x_t^i, y_t^i)$ for all data points in the joint and marginal samples
  $$\mathcal{L} \leftarrow \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{T}\sum_{t=1}^{T}\theta_{\mathbf{x}^i,\mathbf{y}^i}(x_t^i, y_t^i) - \log\left(\frac{1}{T}\sum_{t=1}^{T}\exp(\theta_{\mathbf{x}^i,\mathbf{y}^i}(x_t^i, \tilde{y}_t^i))\right)\right\}.$$
  Do gradient ascent for $\mathcal{L}$
until convergence.

sian components defined as: $\mathrm{p}(z) = \sum_{i=1}^{K}\pi_i\mathcal{N}(z|\mu_i, \Sigma_i)$, where $\mathrm{p}(z)$ is the probability density function (PDF) of the GMM, K is the total number of components in the mixture, π_i denotes the weight of the i-th component satisfying $\sum_{i=1}^{K}\pi_i = 1$, and $\mathcal{N}(z|\mu_i, \Sigma_i)$ is the PDF of a Gaussian with mean μ_i and covariance Σ_i. By varying the parameters K, π_i, μ_i, and Σ_i, a GMM can faithfully approximate an arbitrary distribution. With this, we propose that sampling from GMMs allows us to synthesize arbitrarily complex distributions so that the trained InfoNet can generalize to real-world ones (in a similar spirit to Cranmer et al. (2020); Lavin et al. (2021)).

In our experiments, we set the maximum number of components to 20 to ensure enough diversity in the sampled GMMs. Specifically, we first randomly choose a number K from $\{1, 2, ..., 20\}$, and then perform another sampling of the component weights $\{\pi_i\}_{i=1}^{K}$ such that their sum is one. For each GMM component, we randomly sample its mean from the interval $[-5, 5]$. To generate the covariance matrix, we begin by creating a matrix $\mathbf{D}$ where each element is sampled from the range $[-3, 3]$. Then, the covariance matrix is derived by $\Sigma = \mathbf{D}\mathbf{D}^T + \epsilon\mathbf{I}$, where $\epsilon = 0.01$ is to enforce the matrix to be positive definite. To this end, a random GMM distribution is instantiated, and we can sample from it to get a joint sequence by partitioning z into two parts. Examples of the GMM samplings can be found in Appendix A.9. A training batch contains 32 randomly generated GMM distributions (sequences) with a sample length of 2000. Also, note that each batch is sampled from a different set of GMMs to ensure the training data for InfoNet is diverse and can explore the whole GMM family. Trained with randomly sampled distributions, our model is empowered to estimate MI for an untrained distribution encountered during test time. Please refer to Algorithm 1 for the full training pipeline of the proposed method.

5. Experiments

We concentrate on three aspects related to the training effectiveness and estimation efficiency: 1) Establishing evaluation criteria and collecting data for the proposed InfoNet and baseline methods; 2) Validating and comparing the proposed MI estimation pipeline with other baselines in various settings; 3) Conducting experiments on data with real-world statistics to evaluate performance against other baselines in terms of efficiency and generalization.

5.1. Evaluation Data and Metrics

Evaluation sequences (distributions) are generated with the same protocol described in Sec. 4 and the ground-truth (GT) MI is determined by the following: for a single-component GMM (Gaussian), we apply the analytical formula of MI; otherwise, the Monte-Carlo Integration (MCI) method (Shapiro, 2003) is employed to compute the GT by integrating the known density function.

Setups and Metrics. We evaluate our method and others with the following setups:

- **Test-Time Efficiency.** We compare the computational efficiency of the proposed InfoNet with various baseline methods across different distributions and sequence lengths drawn from the GMM family.
- **Sanity Check.** We use the sequences sampled from Gaussian distributions to benchmark against other methods, a commonly adopted evaluation setting in the MI estimation literature (Belghazi et al., 2018b; Piras et al., 2023). The GT MI values are computed with the analytical formula.
- **GMMs with Multiple Components.** We also analyze the mean and variance of the errors in estimated MI. Specifically, we bin the sampled evaluation sequences based on their ground-truth MI values, such as those with a ground-truth MI of around 0.5. We then report the mean and variance of these errors for each bin of different methods.
- **Mutual Correlation Order Accuracy.** Beyond application domains where the exact MI value is critical, most of the time, for decision-making, the more important is the order of mutual correlations between different random variables. For this, we generate an evaluation set of joint distributions consisting of triplets of random variables $\{(\mathbf{x}, \mathbf{y}, \mathbf{y}')\}$, whose ground-truth order is determined by the computed GT MI (i.e., $\mathbb{I}(\mathbf{x}, \mathbf{y}) > \mathbb{I}(\mathbf{x}, \mathbf{y}')$). We test

Table 1: Comparison of test-time efficiency on GMM distributions with varying lengths (unit: seconds).

SEQ. LENGTH	200	500	1000	2000	5000
KSG-1	0.009	0.024	0.049	0.098	0.249
KSG-5	0.010	0.025	0.049	0.102	0.253
KDE	0.004	0.021	0.083	0.32	1.801
MINE-2000	3.350	3.455	3.607	3.930	4.157
MINE-500	0.821	0.864	0.908	0.991	1.235
MINE-10	0.017	0.017	0.019	0.021	0.027
OURS-1	0.010	0.010	0.011	0.011	0.013
OURS-16	**0.001**	**0.002**	**0.002**	**0.002**	**0.003**

different methods on the triplets to check the correlation order accuracy averaged over all triplets.

- **High Dimensional Independence Testing.** We further employ the slicing technique proposed in sliced mutual information (SMI) (Goldfeld & Greenewald, 2021) for estimating MI between high-dimensional variables with InfoNet. Details are in Appendix A.2. Note that the slicing technique causes minor computational overhead due to parallelization. Due to the lack of GT values, we assess InfoNet's capability to accurately determine the independence between two random vectors.

- **Evaluation with Motion Data.** We verify the generalization of the trained InfoNet on motion data with real-world statistics (e.g., (Radford et al., 2021; Zheng et al., 2023)), where the goal is to check whether the points coming from the same object in motion can be grouped correctly by the estimated MI.

5.2. Results and Comparisons

We report the results and comparisons with three primary baselines. These baselines include: KSG (Kraskov et al., 2004), calculating MI by averaging k-nearest neighbor distances for entropy estimates; KDE (Silverman, 2018), which uses kernel functions to estimate joint and marginal densities, followed by MI computation through integration; and MINE (Belghazi et al., 2018a), employing a similar dual formulation for MI estimation as InfoNet but resorts to optimizing a network for different distributions. All evaluations are conducted on an RTX 4090 GPU and an AMD Ryzen Threadripper PRO 5975WX 32-Core CPU.

Test-Time Efficiency. We compare the time complexity of InfoNet with baseline methods on *new* distributions sampled with varying sequence lengths. The run times, averaged over 100 trials, are presented in Tab. 1, illustrating the efficiency of different approaches. For MINE, the parameters for test-time optimization are: a batch size of 100 and a learning rate of 0.001, while MINE-500 indicates 500 training iterations.

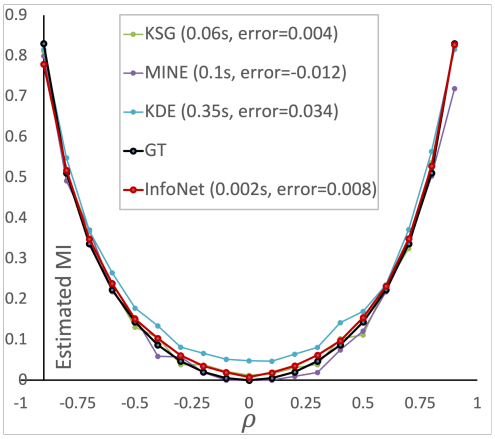


Figure 4: Comparison of MI estimates under Gaussian settings (runtime included).

For our approach, InfoNet-16 denotes the simultaneous estimation of 16 distributions in the batch mode. No training in InfoNet is needed. The results demonstrate that InfoNet significantly outperforms other methods in processing speed for all sequence lengths tested, underscoring our method's efficiency in diverse settings.

Sanity Check on Gaussian. We evaluate the MI estimation accuracy of different methods on Gaussian distributions. In this case, the MI of a (joint) distribution depends on their Pearson correlation coefficient ρ. For a fair comparison, the MINE model is trained with a batch size of 500 for 500 steps at a learning rate of 0.001. The KSG method uses a neighborhood size of 5 for best performance. For each method, the number of data samples from the test distributions is 2000. As shown in Fig. 4, InfoNet predicts MI values closer to the ground-truth MI compared to baseline methods, with mean error shown in the figure's legend. We can see that InfoNet quantitatively achieves a similar error with KSG but is 30$\times$ faster. When compared to MINE, InfoNet runs 50$\times$ faster, while achieving a 30% improvement in accuracy. This sanity check verifies that the proposed InfoNet has an optimal efficiency-accuracy tradeoff than others. More results can be found in Appendix A.4.

GMMs with Multiple Components. We also evaluate the above MI estimators on GMMs with multiple components, which is a more challenging but practical task. Our test dataset is generated as follows: we define 10 MI levels, ranging from 0.0 to 0.9, and create random GMM distributions using the same data generation protocol. The MI of a sampled distribution is calculated using the Monte-Carlo Integration (MCI) method. A GMM distribution is saved to one of the 10 MI levels if its computed MI is within ± 0.02 of that level's value, while also recording the exact MI. The test data generation continues until each MI level has 1,000

Table 2: Error mean and variance of different MI estimators. Methods that do not rely on neural networks are highlighted in Blue, and those leveraging neural networks are colored Green. Numbers highlighted in bold represent the optimal performance achieved by neural estimators.

	MI	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Mean	KSG	0.001	0.001	0.004	0.006	0.008	0.009	0.012	0.015	0.016	0.014
	KDE	0.005	0.010	-0.003	-0.350	-0.071	-0.109	-0.155	-0.199	-0.239	-0.292
	MINE-500	**-0.003**	-0.058	-0.116	-0.173	-0.228	-0.294	-0.344	-0.399	-0.431	-0.485
	MINE-100	-0.008	-0.092	-0.173	-0.251	-0.336	-0.420	-0.504	-0.584	-0.658	-0.742
	InfoNet	0.010	**0.004**	**0.008**	**-0.024**	**-0.040**	**-0.063**	**-0.082**	**-0.101**	**-0.124**	**-0.138**
Variance	KSG	2e-4	3e-4	4e-4	5e-4	6e-4	8e-4	9e-4	9e-4	1e-3	1e-3
	KDE	0.010	0.005	0.001	0.003	0.004	0.005	0.010	0.012	0.014	0.019
	MINE-500	4e-5	0.001	0.004	0.008	0.013	0.018	0.027	0.039	0.052	0.060
	MINE-100	4e-5	5e-4	0.002	0.005	0.009	0.012	0.017	0.025	0.033	0.040
	InfoNet	**1e-5**	**1e-4**	**3e-4**	**8e-4**	**0.001**	**0.002**	**0.004**	**0.005**	**0.007**	**0.009**

test distributions. For each GMM distribution, we sample sequences with a length equal to 2000. These sequences' MI is then estimated using various methods. The mean error and variance for each method across different MI levels are summarized in Tab. 2.

Accordingly, we can make the following observations: 1) Although traditional methods, e.g., KSG, perform relatively well in terms of mean error and variance, they cannot utilize neural networks for computational efficiency. 2) Among the neural methods (InfoNet and variants of MINE), our model achieves much smaller mean errors, and the prediction is more stable (in terms of variance) than MINE, which performs 100 and 500 gradient steps during the test-time training for different distributions. The runtime for MINE-100 and MINE-500 are 0.17 and 0.991 seconds, respectively, while the runtime for InfoNet is 0.011 seconds.

Mutual Correlation Order Accuracy. Now we report the results of various methods for the task of correlation order prediction with varying GMM components (e.g., K ranges from 1 to 10), investigating how order accuracy changes with the difficulty of MI estimation. As outlined in Sec. 5.1, we evaluate the estimated order of 2000 triplets per category (K) against the ground truth established by the MCI method. An order is deemed accurate if the estimated relationship ($\mathbb{I}(\mathbf{x},\mathbf{y}) > \mathbb{I}(\mathbf{x},\mathbf{y}')$ or $\mathbb{I}(\mathbf{x},\mathbf{y}) \leq \mathbb{I}(\mathbf{x},\mathbf{y}')$) aligns with the ground truth. The results are summarized in Tab. 3.

We can see that InfoNet consistently achieves higher order accuracy than the test-time optimization method (MINE), despite both utilizing neural networks. Furthermore, even as the difficulty of MI estimation increases (K from 1 to 10), InfoNet reliably produces accurate order estimates between variables under different joint distributions. This underscores InfoNet's generalization as a neural estimator of correlation order for decision-making. Performance curves comparing InfoNet and GT under different Gaussian components are shown in Appendix A.3 (Fig. 7).

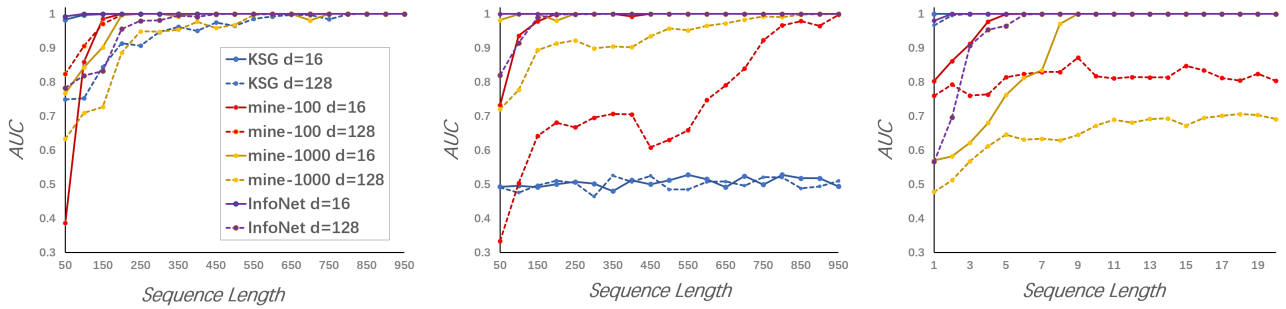
High Dimensional Independence Testing. We assess InfoNet in dealing with high-dimensional random variables (distributions) by the independence test proposed in Sec. 5.1. Even though trained with low-dimensional variables, we can easily adapt InfoNet for high-dimensional data by utilizing the slicing technique proposed in Goldfeld & Greenewald (2021). Specifically, the sliced mutual information (SMI) $\mathbb{SI}(X,Y)$ computed of high-dimensional X and Y with MI estimators (for low-dimensional variables) guarantees that $\mathbb{SI}(X,Y) = 0$ implies $\mathbb{I}(X,Y) = 0$, i.e., independence between X and Y.

Fig. 5 shows the results of the proposed independence testing in three settings. We report the area under the curve (AUC) of the receiver operating characteristic (ROC) for the slicing-empowered InfoNet, MINE, and KSG. The computation of SMI with InfoNet involves 1000 random projection steps in parallel. We obtain the high-dimensional test data with three types of data correlations (Appendix A.5). Each number on the curve is an average over ten trials. For each trial, the independence is evaluated on 100 pairs of RVs with different MI estimators, within which 50 are independent and the remaining 50 are dependent, balancing the labels. More details on how to generate high-dimensional joint distributions can be found in Appendix A.5. To get the plot, we vary the sequence length n and the RV dimension d from 16 to 128, showing the variations of AUC of different methods. The plots in Fig. 5 demonstrate that InfoNet performs better than the other baselines in independence testing of high-dimensional random variables, verifying the effectiveness of InfoNet for dealing with high-dimensional data, especially with short sequences. More results on high-dimensional data can be found in Appendix A.6.

Validation on Out-of-Domain Motion Data. We further evaluate our model's generalization to out-of-domain data. Specifically, we leverage InfoNet to perform mutual information estimation on motion trajectories of pixels in a video,

Table 3: Correlation order prediction accuracy of different MI estimators. Methods without neural networks are highlighted in Blue, and neural estimators are colored Green. Performance is reported with various numbers of components in GMMs.

NO. OF COMPS.	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
KSG	98.7	99.0	98.2	98.0	97.9	97.7	97.6	97.5	97.0	97.3
KDE	97.4	97.7	97.9	97.5	97.9	97.8	97.0	97.4	97.4	97.4
MINE-500	98.5	91.2	90.8	87.2	84.5	83.7	81.2	79.6	81.3	78.1
MINE-100	94.6	77.1	75.4	71.6	67.5	69.4	66.5	66.3	68.7	66.4
MINE-10	60.9	56.1	55.1	54.3	52.4	54.9	53.7	50.4	53.1	52.5
INFONET	**99.8**	**99.5**	**99.0**	**99.2**	**99.1**	**99.2**	**99.0**	**99.2**	**99.3**	**99.5**



(a) Correlation Type I (b) Correlation Type II (c) Correlation Type III

Figure 5: Independence testing under three types of data correlations. Each curve in the plots depicts the area under the curve (AUC) of the receiver operating characteristic (ROC) with respect to sequence length n. Four MI estimators are compared: InfoNet, KSG, MINE-100, and MINE-1000 (i.e., MINE trained with 100 and 1000 gradient steps during test-time optimization), each with two dimensions (16 and 128). The curves obtained by InfoNet (with the slicing technique) are constantly higher than the others, which demonstrates the effectiveness of InfoNet for dealing with high-dimensional data.

and then use the estimated MI to perform object segmentation by thresholding the MIs. If the estimation is correct, pixels from the same object should be grouped together.

We use the Pointodyssey dataset (Zheng et al., 2023), consisting of long videos that provide rich ground-truth trajectories. Given a pixel P_i in the first frame of a video, we can extract the locations of the corresponding pixels in all frames with the ground-truth trajectories. Then we utilize InfoNet to compute mutual information between the locations of any two pixels P_i and P_i in the first frame. Since InfoNet does not resort to test-time optimization, the computation is efficient, e.g., less than 5 seconds for hundreds of pairs. Instead of picking one threshold γ for the grouping, we vary it in an increment of 0.01 to plot the precision-recall (P-R) curves with the help of the ground-truth object masks. Fig. 6 shows the P-R curves obtained for each MI estimation method on Pointodyssey. For more details, please refer to Appendix A.7. Results in Fig. 6 demonstrate InfoNet's ability to achieve higher segmentation performance with sound generalization to out-of-domain data.

6. Related Works

MI quantifies the statistical dependence between variables through a variety of nonparametric and parametric approaches. Nonparametric methods, such as K-Nearest Neighbors (KNN) and Kernel Density Estimation (KDE), estimate MI without assuming specific probability distributions (Reshef et al., 2011; Kinney & Atwal, 2014; Khan et al., 2007; Kwak & Choi, 2002; Kraskov et al., 2004; Pál et al., 2010; Gao et al., 2015b; 2017; Runge, 2018; Lord et al., 2018; Moon et al., 1995; Steuer et al., 2002; Gretton et al., 2005; Kumar et al., 2021). These methods, however, suffer from limitations such as sensitivity to parameter choice, curse of dimensionality, computational complexity, and assumptions about continuity (Suzuki et al., 2008; Walters-Williams & Li, 2009; Gao et al., 2018; Mukherjee et al., 2020; Fukumizu et al., 2007; Estévez et al., 2009; Bach, 2022). Binning methods and adaptive partitioning offer nonparametric alternatives but are constrained by bin/partition selection and the curse of dimensionality (Lugosi & Nobel, 1996; Darbellay & Vajda, 1999; Cellucci et al., 2005; Fernando et al., 2009; Cakir et al., 2019; Marx et al., 2021; Thévenaz & Unser, 2000; Paninski, 2003; Knops et al., 2006; Tsimpiris et al., 2012). On the

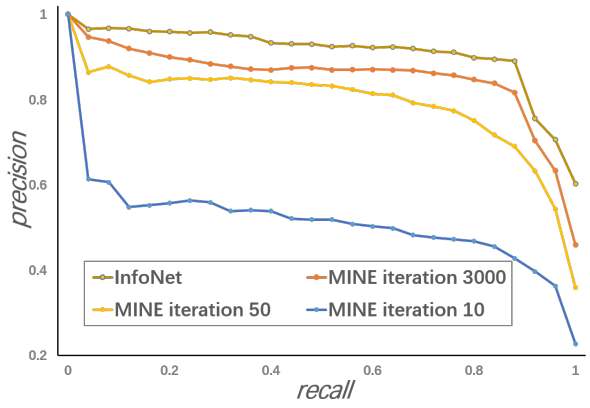


Figure 6: Precision-Recall curves of MI-based segmentation on Pointodyssey, verifying MI estimators' generalization to out-of-domain data.

other hand, parametric methods assume specific distributions, such as Gaussian, but their accuracy is contingent upon correct assumptions and parameter estimation (Hulle, 2005; Gupta & Srivastava, 2010; Sugiyama et al., 2012; Gao et al., 2015a; Ince et al., 2017; Suzuki et al., 2008; Walters-Williams & Li, 2009).

Measuring and optimizing MI with limited sample sizes presents a challenge (Treves & Panzeri, 1995; McAllester & Stratos, 2020). Nevertheless, alternative measurements within a Reproducing Kernel Hilbert Space (RKHS) have demonstrated effectiveness in detecting statistical dependence (Gretton et al., 2005). Singular Value Decomposition (SVD) (Anantharam et al., 2013; Makur et al., 2015), Alternating Conditional Expectation (ACE) algorithm (Breiman & Friedman, 1985; Buja, 1990; Huang & Xu, 2020; Almaraz-Damian et al., 2020), and rank correlation (Kendall, 1938; Klaassen & Wellner, 1997) are widely used conventional methods. Recently, neural network approaches have also been proposed (Xu & Huang, 2020).

Numerous works have addressed the scalable computation of MI and statistical dependences (Lopez-Paz et al., 2013; Mary et al., 2019; Goldfeld & Greenewald, 2021; Chen et al., 2022). Our proposed InfoNet offers an orthogonal alternative to these methods. Instead of striving for a more accurate approximation of the highly nonlinear MI or devising advanced yet computationally friendly correlation metrics, InfoNet focuses on MI estimation by encoding the optimization of its objectives into neural networks through pertaining, bypassing test-time optimization and conceptually allows for more efficient and accurate solutions to these complex correlation measures. The proposed method is also related to simulation-based intelligence (Cranmer et al., 2020; Ramon et al., 2021).

7. Discussion

We present InfoNet, a novel neural network architecture for efficient MI estimation. Utilizing the attention mechanism and large-scale training, our approach circumvents time-consuming test-time optimization and demonstrates generalization capabilities. We extensively evaluated InfoNet's effectiveness on various distribution families and applications, emphasizing its efficiency-accuracy trade-off and order-preserving properties. We validate InfoNet's potential in the fields requiring real-time MI estimation and expect that our work can facilitate further exploration of neuralizing the computation of MI and other information-theoretic quantities. We also expect the proposed method and trained models can benefit applications that require estimating a vast amount of correlation in a low time budget. Future work could investigate leveraging the proposed training scheme for directly estimating the mutual information between high-dimensional random variables.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgement:

This work is supported by the HKU-100 Award, a donation from the Musketeers Foundation, and the HKU Seed Fund for Basic Research # 2202100553.

References

Almaraz-Damian, J.-A., Ponomaryov, V., Sadovnychiy, S., and Castillejos-Fernandez, H. Melanoma and nevus skin lesion classification using handcraft and deep learning feature fusion via mutual information measures. *Entropy*, 22(4):484, 2020.

Anantharam, V., Gohari, A., Kamath, S., and Nair, C. On maximal correlation, hypercontractivity, and the data processing inequality studied by erkip and cover. *arXiv preprint arXiv:1304.6133*, 2013.

Bach, F. Information theory with kernel methods. *IEEE Transactions on Information Theory*, 69(2):752–775, 2022.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018a.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Ben-

gio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018b.

Breiman, L. and Friedman, J. H. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598, 1985.

Buja, A. Remarks on functional canonical variates, alternating least squares methods and ace. *The Annals of Statistics*, pp. 1032–1069, 1990.

Cakir, F., He, K., Bargal, S. A., and Sclaroff, S. Hashing with mutual information. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2424–2437, 2019.

Cellucci, C. J., Albano, A. M., and Rapp, P. E. Statistical validation of mutual information calculations: Comparison of alternative numerical algorithms. *Physical review E*, 71(6):066208, 2005.

Chen, Y., Li, Y., Weller, A., et al. Scalable infomin learning. *Advances in Neural Information Processing Systems*, 35: 2226–2239, 2022.

Cranmer, K., Brehmer, J., and Louppe, G. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Czyż, P., Grabowski, F., Vogt, J. E., Beerenwinkel, N., and Marx, A. Beyond normal: On the evaluation of mutual information estimators. *arXiv preprint arXiv:2306.11078*, 2023.

Darbellay, G. A. and Vajda, I. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, 1999.

Donsker, M. D. and Varadhan, S. S. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36 (2):183–212, 1983.

Durante, F. and Sempi, C. Copula theory: an introduction. In *Copula Theory and Its Applications: Proceedings of the Workshop Held in Warsaw, 25-26 September 2009*, pp. 3–31. Springer, 2010.

Estévez, P. A., Tesmer, M., Perez, C. A., and Zurada, J. M. Normalized mutual information feature selection. *IEEE Transactions on neural networks*, 20(2):189–201, 2009.

Fernando, T., Maier, H., and Dandy, G. Selection of input variables for data driven models: An average shifted histogram partial mutual information estimator approach. *Journal of Hydrology*, 367(3-4):165–176, 2009.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. *Advances in neural information processing systems*, 20, 2007.

Gao, S., Steeg, G. V., and Galstyan, A. Estimating mutual information by local gaussian approximation. *arXiv preprint arXiv:1508.00536*, 2015a.

Gao, S., Ver Steeg, G., and Galstyan, A. Efficient estimation of mutual information for strongly dependent variables. In *Artificial intelligence and statistics*, pp. 277–286. PMLR, 2015b.

Gao, W., Kannan, S., Oh, S., and Viswanath, P. Estimating mutual information for discrete-continuous mixtures. *Advances in neural information processing systems*, 30, 2017.

Gao, W., Oh, S., and Viswanath, P. Demystifying fixed k-nearest neighbor information estimators. *IEEE Transactions on Information Theory*, 64(8):5629–5661, 2018.

Goldfeld, Z. and Greenewald, K. Sliced mutual information: A scalable measure of statistical dependence. *Advances in Neural Information Processing Systems*, 34:17567–17578, 2021.

Gretton, A., Smola, A., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. Kernel constrained covariance for dependence measurement. In *International Workshop on Artificial Intelligence and Statistics*, pp. 112–119. PMLR, 2005.

Gupta, M. and Srivastava, S. Parametric bayesian estimation of differential entropy and relative entropy. *Entropy*, 12 (4):818–843, 2010.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Huang, S.-L. and Xu, X. On the sample complexity of hgr maximal correlation functions for large datasets. *IEEE Transactions on Information Theory*, 67(3):1951–1980, 2020.

Hulle, M. M. V. Edgeworth approximation of multivariate differential entropy. *Neural computation*, 17(9):1903–1910, 2005.

Ince, R. A., Giordano, B. L., Kayser, C., Rousselet, G. A., Gross, J., and Schyns, P. G. A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. *Human brain mapping*, 38(3):1541–1573, 2017.

Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

Kendall, M. G. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Khan, S., Bandyopadhyay, S., Ganguly, A. R., Saigal, S., Erickson III, D. J., Protopopescu, V., and Ostrouchov, G. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Physical Review E*, 76(2):026209, 2007.

Kinney, J. B. and Atwal, G. S. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9): 3354–3359, 2014.

Klaassen, C. A. and Wellner, J. A. Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli*, pp. 55–77, 1997.

Knops, Z. F., Maintz, J. A., Viergever, M. A., and Pluim, J. P. Normalized mutual information based registration using k-means clustering and shading correction. *Medical image analysis*, 10(3):432–439, 2006.

Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

Kumar, A., Zhou, Y., and Xiang, J. Optimization of vmd using kernel-based mutual information for the extraction of weak features to detect bearing defects. *Measurement*, 168:108402, 2021.

Kwak, N. and Choi, C.-H. Input feature selection by mutual information based on parzen window. *IEEE transactions on pattern analysis and machine intelligence*, 24(12): 1667–1671, 2002.

Lavin, A., Krakauer, D., Zenil, H., Gottschlich, J., Mattson, T., Brehmer, J., Anandkumar, A., Choudry, S., Rocki, K., Baydin, A. G., et al. Simulation intelligence: Towards a new generation of scientific methods. *arXiv preprint arXiv:2112.03235*, 2021.

Lopez-Paz, D., Hennig, P., and Schölkopf, B. The randomized dependence coefficient. *Advances in neural information processing systems*, 26, 2013.

Lord, W. M., Sun, J., and Bollt, E. M. Geometric k-nearest neighbor estimation of entropy and mutual information. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(3), 2018.

Lugosi, G. and Nobel, A. Consistency of data-driven histogram methods for density estimation and classification. *The Annals of Statistics*, 24(2):687–706, 1996.

Makur, A., Kozynski, F., Huang, S.-L., and Zheng, L. An efficient algorithm for information decomposition and extraction. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 972–979. IEEE, 2015.

Marx, A., Yang, L., and van Leeuwen, M. Estimating conditional mutual information for discrete-continuous mixtures using multi-dimensional adaptive histograms. In *Proceedings of the 2021 SIAM international conference on data mining (SDM)*, pp. 387–395. SIAM, 2021.

Mary, J., Calauzenes, C., and El Karoui, N. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pp. 4382–4391. PMLR, 2019.

McAllester, D. and Stratos, K. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pp. 875–884. PMLR, 2020.

Moon, Y.-I., Rajagopalan, B., and Lall, U. Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3):2318, 1995.

Mukherjee, S., Asnani, H., and Kannan, S. Ccmi: Classifier based conditional mutual information estimation. In *Uncertainty in artificial intelligence*, pp. 1083–1093. PMLR, 2020.

Pál, D., Póczos, B., and Szepesvári, C. Estimation of rényi entropy and mutual information based on generalized nearest-neighbor graphs. *Advances in Neural Information Processing Systems*, 23, 2010.

Paninski, L. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

Piras, D., Peiris, H. V., Pontzen, A., Lucie-Smith, L., Guo, N., and Nord, B. A robust estimator of mutual information for deep learning interpretability. *Machine Learning: Science and Technology*, 4(2):025006, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ramon, E., Triginer, G., Escur, J., Pumarola, A., Garcia, J., Giro-i Nieto, X., and Moreno-Noguer, F. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5620–5629, 2021.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

Reynolds, D. A. et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

Rhodes, B., Xu, K., and Gutmann, M. U. Telescoping density-ratio estimation. *Advances in neural information processing systems*, 33:4905–4916, 2020.

Runge, J. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, pp. 938–947. PMLR, 2018.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Shapiro, A. Monte carlo sampling methods. *Handbooks in operations research and management science*, 10:353–425, 2003.

Silverman, B. W. *Density estimation for statistics and data analysis*. Routledge, 2018.

Steuer, R., Kurths, J., Daub, C. O., Weise, J., and Selbig, J. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 18(suppl_2): S231–S240, 2002.

Sugiyama, M., Suzuki, T., and Kanamori, T. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Suzuki, T., Sugiyama, M., Sese, J., and Kanamori, T. Approximating mutual information by maximum likelihood density ratio estimation. In *New challenges for feature selection in data mining and knowledge discovery*, pp. 5–20. PMLR, 2008.

Thévenaz, P. and Unser, M. Optimization of mutual information for multiresolution image registration. *IEEE transactions on image processing*, 9(12):2083–2099, 2000.

Treves, A. and Panzeri, S. The upward bias in measures of information derived from limited data samples. *Neural Computation*, 7(2):399–407, 1995.

Tsimpiris, A., Vlachos, I., and Kugiumtzis, D. Nearest neighbor estimate of conditional mutual information in feature selection. *Expert Systems with Applications*, 39 (16):12697–12708, 2012.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Walters-Williams, J. and Li, Y. Estimation of mutual information: A survey. In *Rough Sets and Knowledge Technology: 4th International Conference, RSKT 2009, Gold Coast, Australia, July 14-16, 2009. Proceedings 4*, pp. 389–396. Springer, 2009.

Xu, X. and Huang, S.-L. Maximal correlation regression. *IEEE Access*, 8:26591–26601, 2020.

Zheng, Y., Harley, A. W., Shen, B., Wetzstein, G., and Guibas, L. J. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. *arXiv preprint arXiv:2307.15055*, 2023.

A. Appendix

A.1. Copulas

To enhance the efficiency of MI estimation, we introduce the method called copula Durante & Sempi (2010) during the data preprocessing stage. This approach is initiated based on a fundamental property of MI: given that $f, g : \mathbb{R} \to \mathbb{R}$ are arbitrary strictly increasing functions, the following equation holds true:

$$\mathbb{I}\left(f(X), g(Y)\right) = \mathbb{I}\left(X, Y\right). \tag{3}$$

Specifically, drawing inspiration from (Pál et al., 2010), we choose mappings $f = F_X$ and $g = F_Y$, representing the cumulative distribution functions (CDFs) of random variables X and Y. For continuous F_X and F_Y, the marginal distribution uniformly spans $[0, 1]$.

While the specific CDF of X and Y is not known in our situations, we employ the empirical CDF $\left(\widehat{F}_X, \widehat{F}_Y\right)$ as an alternative. Given a sequence $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ with length n, where each sample X_i, $i = 1, \cdots, n$, originates from an unknown distribution, the empirical CDF is defined as follows:

$$\widehat{F}_X(x) = \frac{1}{n} \operatorname{card}\left(\{i : 1 \leq i \leq n, x \leq X_i\}\right), \quad x \in \mathbb{R}, \tag{4}$$

where $\operatorname{card}(\cdot)$ denotes the cardinality of the set. Note that while $\widehat{F}_X$ does not establish a bijection between $\mathbb{R}$ and the interval $[0, 1]$, it is quite straightforward to create a bijection through interpolation while preserving the order of the sampling points. This ensures the invariance of MI, which remains unaffected by the transformation.

Our approach involves using the empirical CDF of X and Y to map them to a uniform distribution between $[0, 1]$ prior to training and evaluation. In practice, this mapping process can be reduced to a simple sorting step:

$$f(x) = \frac{1}{n} \operatorname{card}\left(\{i : 1 \leq i \leq n, x \leq X_i\}\right), \quad x = X_1, X_2, \cdots, X_n, \tag{5}$$

and

$$g(y) = \frac{1}{n} \operatorname{card}\left(\{i : 1 \leq i \leq n, y \leq Y_i\}\right), \quad y = Y_1, Y_2, \cdots, Y_n, \tag{6}$$

which are strictly increasing mappings that satisfy the requirement stated in equation 3.

A.2. Sliced Mutual Information

To estimate high-dimensional MI, we adopt the SMI concept (Goldfeld & Greenewald, 2021), which averages the MI between one-dimensional random projections of variables. Let X and Y be random variables with dimensions d_x and d_y respectively. SMI is thus the expected MI across these one-dimensional projections

$$\text{SMI}(X; Y) = \mathbb{E}_{\phi,\psi}\left[\mathbb{I}(\phi(X); \psi(Y))\right] = \frac{1}{S_{d_x-1} S_{d_y-1}} \oint_{\mathbb{S}^{d_x-1}} \oint_{\mathbb{S}^{d_y-1}} \mathbb{I}\left(\theta^\top X; \phi^\top Y\right) \mathrm{d}\theta \mathrm{d}\phi \tag{7}$$

Here, $\mathbb{S}^{d-1}$ denotes the d-dimensional sphere (whose surface area is designated by S_{d-1}), ϕ and ψ are vectors used for linear projection from high-dimensional space to one-dimensional space, and $\mathbb{E}_{\phi,\psi}$ denotes the expectation over these projection functions.

While SMI typically yields lower values compared to MI, it retains many of the intrinsic properties of MI and exhibits a certain degree of correlation with it. This inter-connectedness is crucial, as it implies that while SMI offers a novel approach to handling high-dimensional data, it still adheres to the fundamental principles of MI, thereby ensuring consistency in its theoretical foundations and practical applications.

A.3. Additional Results on GMMs with Multiple Components

We demonstrate the capability of the InfoNet model in handling GMMs that comprise 1 to 10 Gaussian components. We evaluate the model's precision in estimating MI values by comparing these estimates with ground truth values. The results, depicted in Figure 7, reveal that the InfoNet model accurately estimates MI on Gaussian mixture distributions, yielding estimates that are in close agreement with the ground truth values.

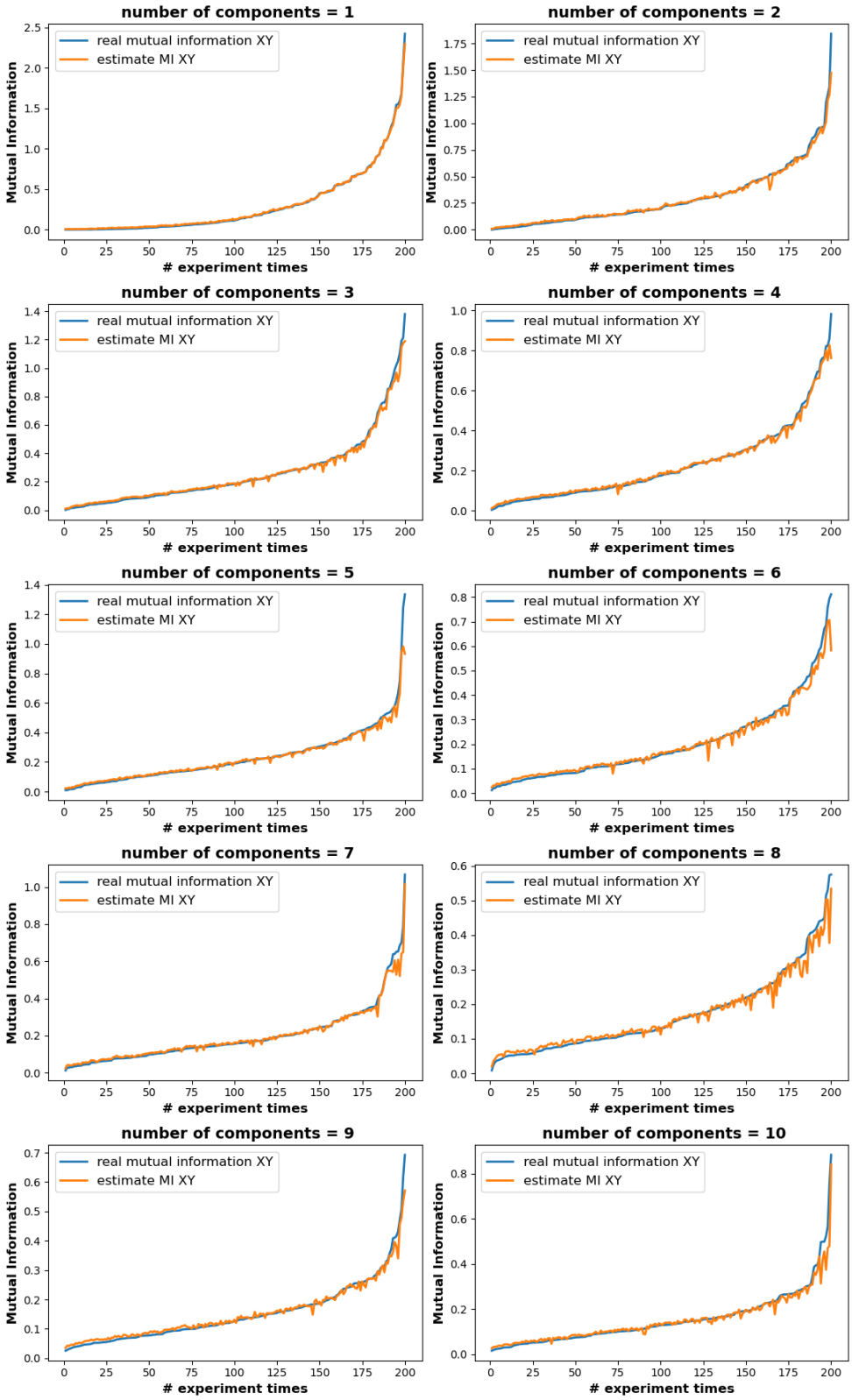


Figure 7: Performance of our InfoNet model across various numbers of Gaussian components. The assessment is based on 200 randomly generated joint distributions in each category (number K of GMM components), then sorted according to the value of ground-truth mutual information.

A.4. Performance on Other Distributions

Paper (Czyż et al., 2023) provides a diverse family of distributions with known ground truth mutual information. We select three one-dimension distributions to test our InfoNet performance, note that our model has been only trained on GMM distributions and without any additional training.

Half-Cube Map Applying the half-cube homeomorphism $h(x) = |x|^{3/2}\operatorname{sign}(x)$ to Gaussian variables X and Y, this could lengthen the tail. The transformation does not influence the ground truth value of MI.

Asinh Mapping Applying inverse hyperbolic sine function $\operatorname{asinh} x = \log\left(x + \sqrt{1+x^2}\right)$ to shorten the tails, this transformation does not change the ground truth value of MI.

Additive Noise Let independent r.v. $X \sim \text{Uniform}(0,1)$ and $N \sim \text{Uniform}(-\varepsilon, \varepsilon)$, where ε is the noise level. For $Y = X + N$, we could derive $\mathbf{I}(X;Y)$ analytically.

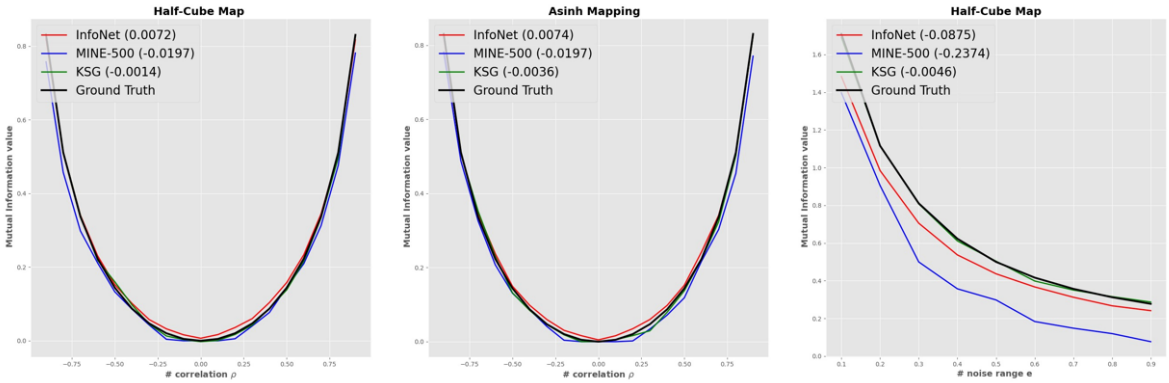


Figure 8: Evaluation of performance on distribution other than GMM, comparing with MINE with 500 training iterations and KSG with nearest neighbor number $k = 1$.

Fig. 8 shows our result on other distributions despite the Mixture of Gaussian distributions. Due to the introduction of the copula, our model can suit different monotonic transformations well and produce good estimations for Half-Cube Map and Asinh Mapping. Also, our model performs well on Additive noise, evidencing good generalization ability as we do not train it on any uniform distributions and additive noise.

A.5. Three types of dependencies between X and Y

Below are three different relationships between X and Y in high dimensional independence test in sec. 5.2.

(a) **One feature (linear)**: $X, Z \sim \mathcal{N}(0, \mathrm{I}_d)$ i.i.d. and $Y = \frac{1}{\sqrt{2}}\left(\frac{1}{\sqrt{d}}\left(\mathbf{1}^\top X\right)\mathbf{1} + Z\right)$, where $\mathbf{1} := (1, \ldots, 1)^\top \in \mathbb{R}^d$.

(b) **Two features**: $X, Z \sim \mathcal{N}(0, \mathrm{I}_d)$ i.i.d. and $Y_i = \frac{1}{\sqrt{2}} \begin{cases} \frac{1}{d}\left(\mathbf{1}_{\lfloor d/2 \rfloor} 0 \ldots 0\right)^\top X + Z_i, & i \le \frac{d}{2} \\ \frac{1}{d}\left(0 \ldots 0 \mathbf{1}_{\lceil d/2 \rceil}\right)^\top X + Z_i, & i > \frac{d}{2}. \end{cases}$

(c) **Independent coordinates**: $X, Z \sim \mathcal{N}(0, \mathrm{I}_d)$ i.i.d. and $Y = \frac{1}{\sqrt{2}}(X + Z)$.

A.6. Additional Result On High Dimension

Similarly, we validate the capability of InfoNet in classifying the correct correlation order on d-dimensional Gauss distributions: $(X, Y) = \left((X^1, X^2, \cdots X^d), (Y^1, Y^2, \cdots Y^d)\right) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

This result shows that our InfoNet model reaches high accuracy and still costs low time complexity. Since our model allows parallel computing on multiple GPUs, it can compute the MI of multiple projected variables in one feed-forward process.

A.7. Results of Validation on Out-of-Domain Motion Data

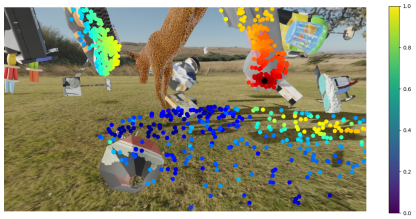
In this section, we provide detailed results of the experiments on motion data.

Table 4: Correlation order accuracy of different MI estimators. Methods that do not rely on neural networks are highlighted in Blue, and those leveraging neural networks are colored Green. MINE-100 means training MINE method for 100 iterations, InfoNet-100 means we do 100 times random projection to get an average.

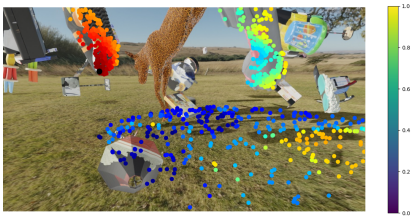
DIMENSIONS	2	3	4	5	6	7	8	9	10
KSG	94.4	95.5	91.8	92	94.1	93.6	94.1	94.1	94.2
ENERGY DISTANCE	49.6	51.2	52.2	51.5	52.5	49.6	48.7	50.2	51.3
MINE-100	78.5	82.1	86.7	84.7	88.4	89.8	90.1	90.4	90
MINE-1000	93.6	93.9	94.4	94.3	91.6	91.7	89.5	91	90.3
MINE-5000	96.2	**97**	**97**	96.2	94.9	94.2	93.2	92.8	93
INFONET-100	93.7	94.6	94.4	95.7	93.3	95.8	95.8	95.4	93.8
INFONET-500	94.9	93.7	95.7	95.8	97.1	96.4	97.2	97.8	96.8
INFONET-1500	**97.7**	96.4	96.2	**97.9**	**97.4**	**98.1**	**98.2**	**97.3**	**98.3**

It is worth noting that certain trajectories provided may contain unreasonable values such as "inf" or "-50000". To address this issue in the dataset, we apply a filtering process to ensure that only points appearing throughout the entire video are considered for analysis.

Fig. 9 and Fig. 10 show the visualization of estimated mutual information between one selected point and other points in the videos. Fig. 11 presents the individual PR curves for each object, while Fig 12 provides the comparison of PR curves across different methods on each object.



(a) Estimated Mutual Information with point in object 1 (highlighted black).



(b) Estimated Mutual Information with point in object 2 (highlighted black).

Figure 9: Visualization results using InfoNet model.

A.8. Additional Validation Results on the Order of Estimated Slice Mutual Information

In this section, we present additional experiments to validate our estimated sliced MI has a strong correlation with the ground truth in multi-dimensional. We use the SpatialMultiOmniglot dataset, following the Rhodes et al. (2020). Our experiment aims to assess the sliced mutual information (MI) between two random variables, u and v, obtained through the following steps:

First, we organize the Omniglot data into alphabets $\{A_i\}_{i=1}^{l}$, with each A_i containing n_i characters, each character represented in 20 variants. Thus, $A_i = \{\{a^i_{j,k}\}_{k=1}^{20}\}_{j=1}^{n_i}$, where $a^i_{j,k}$ is the k-th variant of the j-th character in the i-th alphabet. Sample d indices randomly: $j = (j_1, \ldots, j_d)$ from d different alphabets, where each j_i represents a specific character from one alphabet. Sample two independent, identically distributed (i.i.d.) vectors k and k' from $\prod_{i=1}^{d}$ Uniform(20), representing different versions of the characters:

$$k = (k_1, \ldots, k_d), \quad k' = (k'_1, \ldots, k'_d).$$

Then a datapoint $x = (u, v)$ is defined by:

$$u = (a^1_{j_1,k_1}, \ldots, a^d_{j_d,k_d}), \quad v = (a^1_{j_1+1,k'_1}, \ldots, a^d_{j_d+1,k'_d}).$$

This ensures that u_i and v_i are sequential characters within their alphabet, albeit with randomized versions.

The ground-truth mutual information $I(U, V)$ between u and v can be computed as (see Rhodes et al., 2020):

$$I(U, V) = \sum_{i=1}^{d} \log n_i$$

We present estimation results for dimensions $d = 4$ and $d = 9$ employing the sliced mutual information (SMI) technique. SMI correlates highly with MI and retains many of its properties. As noted in the SMI paper (Goldfeld & Greenewald, 2021), the SMI value tends to be systematically lower than the actual MI value. To compensate for this, we apply a constant scaling to the estimated SMI and compare it with the ground-truth MI values, which helps to verify if the SMI aligns with the MI in terms of order accuracy.

Results for $d = 4$ and $d = 9$ using sliced mutual information are shown below:

Dimension	Method	Value 1	Value 2	Value 3	Value 4	Value 5	Value 6	Value 7
$d = 4$	Ground Truth MI	11.69	12.16	13.21	13.51	14.19	14.28	15.41
	InfoNet estimated SMI	0.164	0.167	0.177	0.178	0.189	0.190	0.199
	SMI $\times$ a constant	12.25	12.47	13.22	13.30	14.12	14.20	14.87
$d = 9$	Ground Truth MI	26.96	29.76	30.01	30.11	30.87	31.14	32.12
	InfoNet estimated SMI	0.061	0.067	0.068	0.068	0.069	0.070	0.072
	SMI $\times$ a constant	27.20	29.88	30.33	30.33	30.78	31.22	32.11

Table 5: Results for $d = 4$ and $d = 9$.

From these results, we observe that the extension of InfoNet to real-world high-dimensional tasks with SMI can accurately capture the correlation order measured by the mutual information. Moreover, with a scaling that compensates for the systematic bias, InfoNet can faithfully capture the mutual information quantity between two high-dimensional random variables.

A.9. Data Distributions

In this section, we provide several plots to visualize the sequences sampled from randomly generated Gaussian mixture distributions used for training.

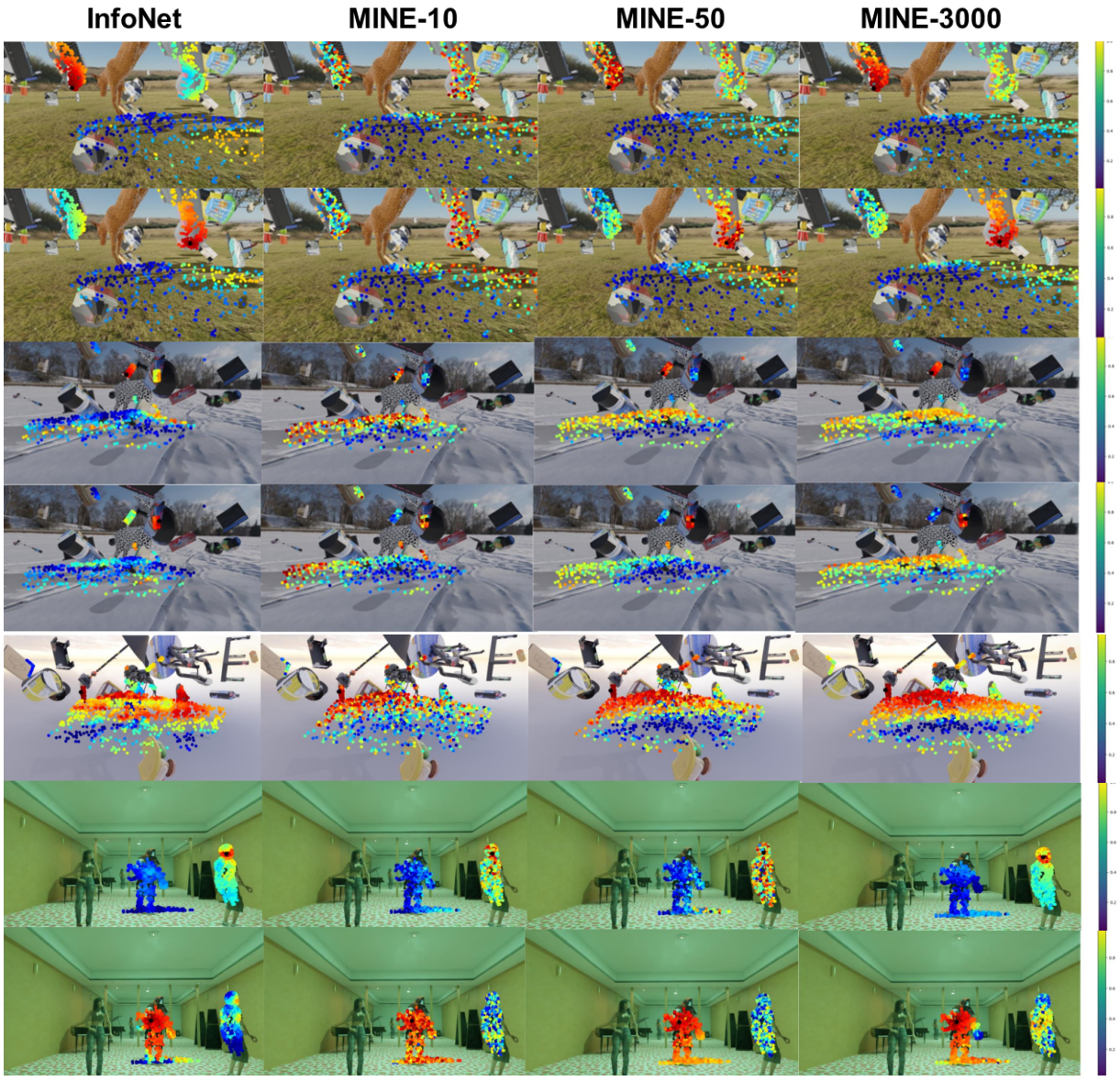
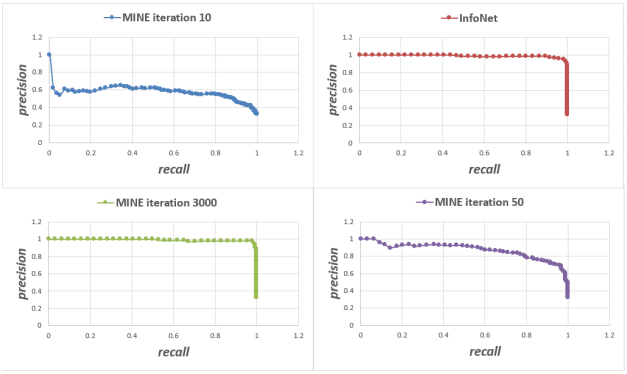
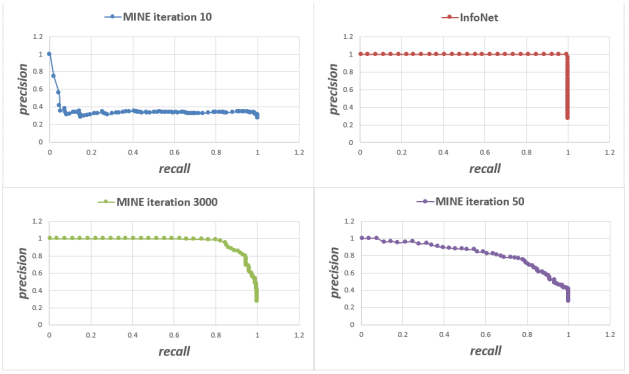


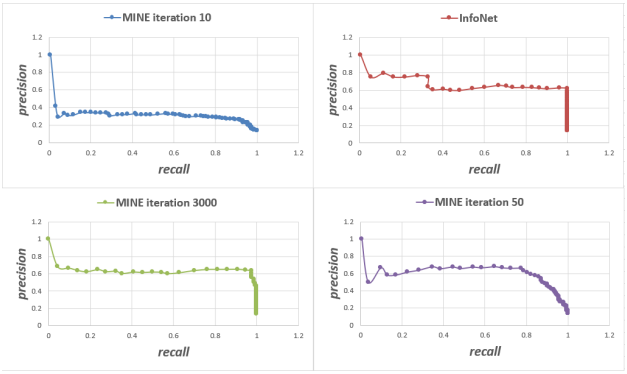
Figure 10: Visual comparison of estimated MI of pixel locations between our model and MINE on the video datasets. Large value in red while small value in blue. From left to right: InfoNet (no test-time optimization), MINE (test-time gradient steps 10), MINE (test-time gradient steps 50), and MINE (test-time gradient steps 3000).



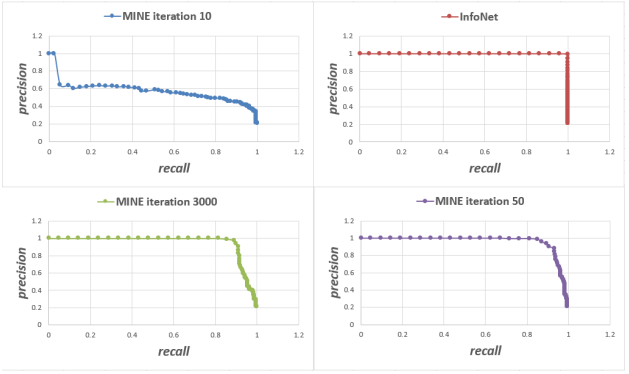
(a)



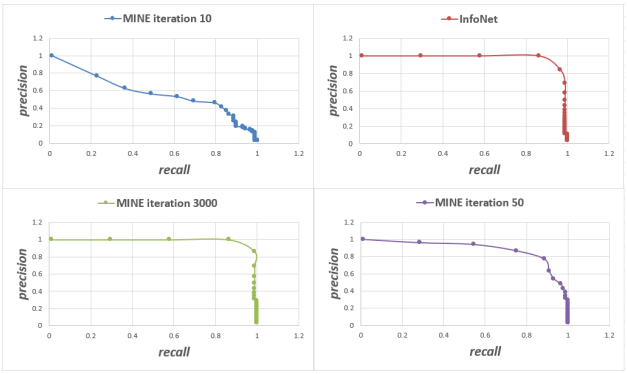
(b)



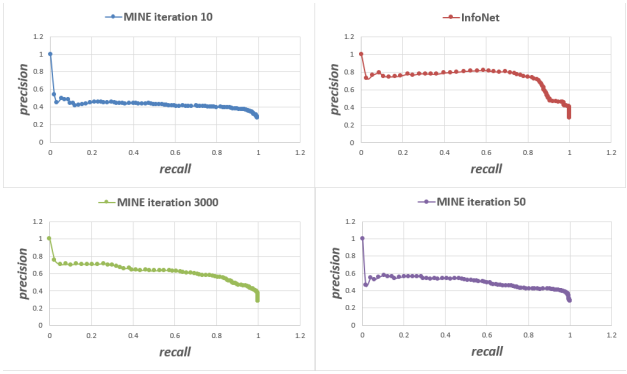
(c)



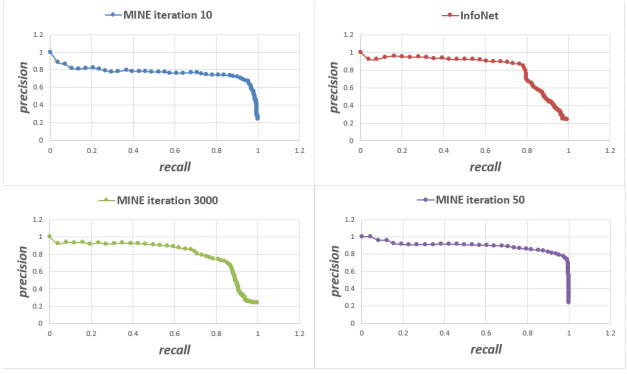
(d)



(e)

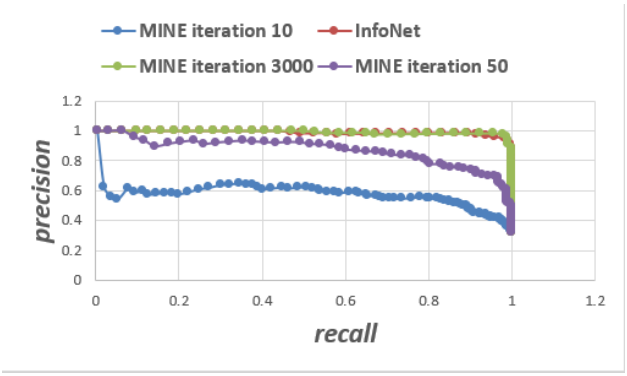


(f)

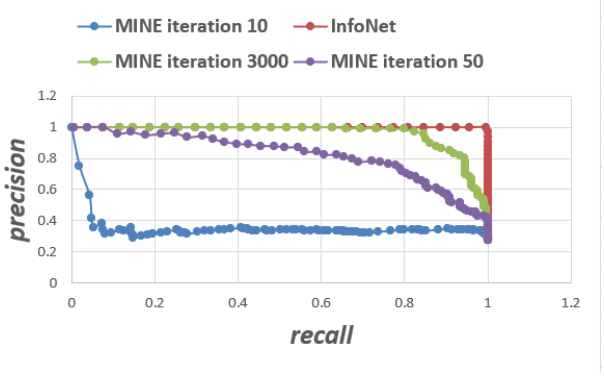


(g)

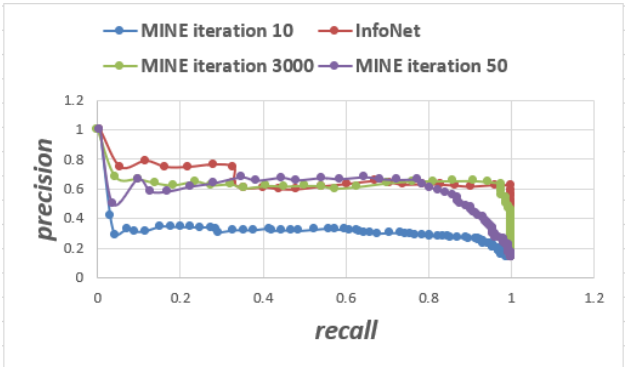
Figure 11: Individual PR graph of our model and MINE. In the experiments conducted on video datasets, InfoNet exhibited notably high stability compared to MINE.



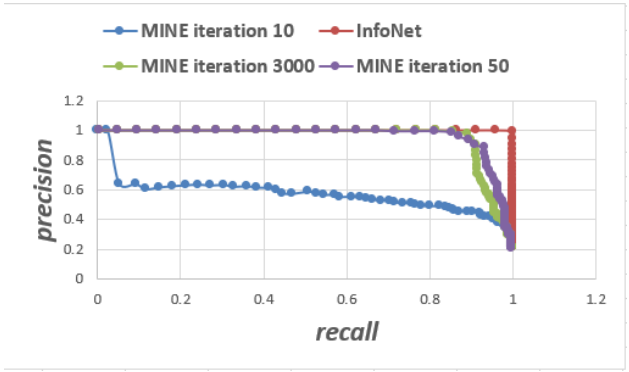
(a)



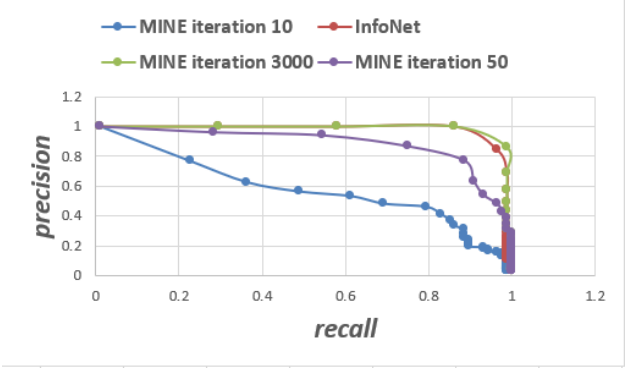
(b)



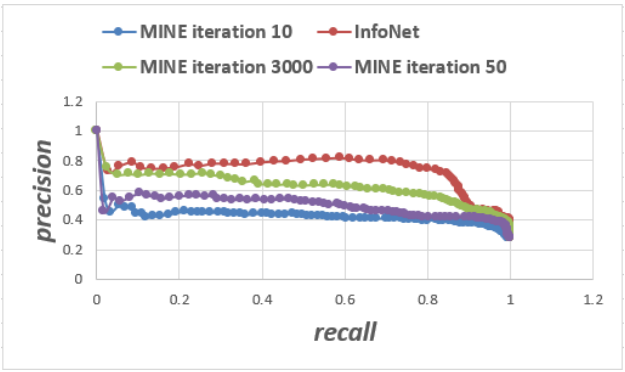
(c)



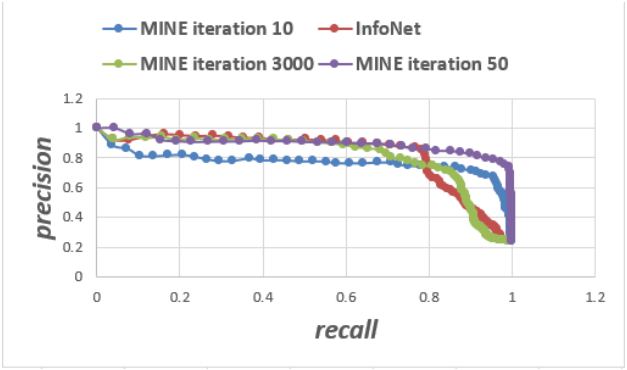
(d)



(e)



(f)



(g)

Figure 12: Comparison between PR graphs of our model and MINE. In the same video dataset, InfoNet consistently exhibits superior performance compared to MINE.

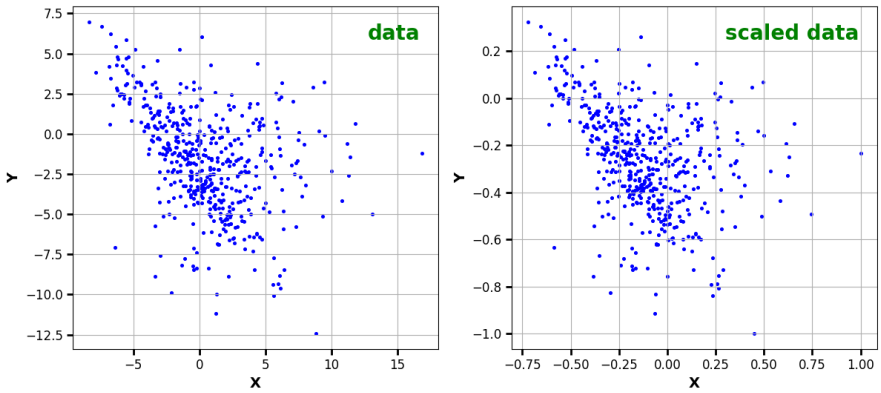


Figure 13: Data points sampled from one mog distribution with 3 components, MI between X and Y is 0.316.

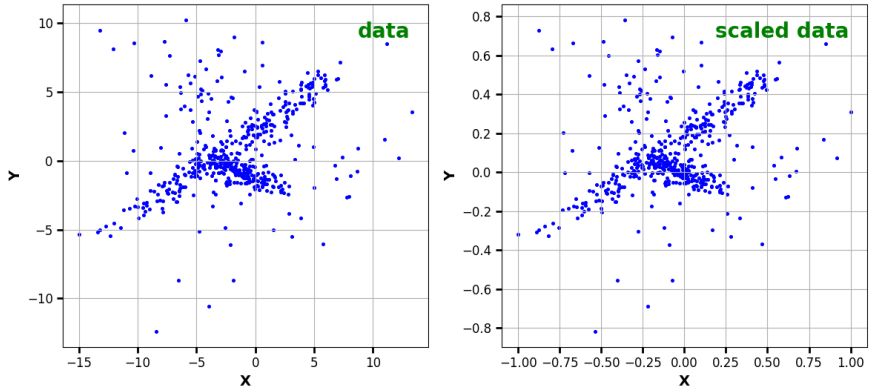


Figure 14: Data points sampled from one mog distribution with 7 components, MI between X and Y is 0.510.

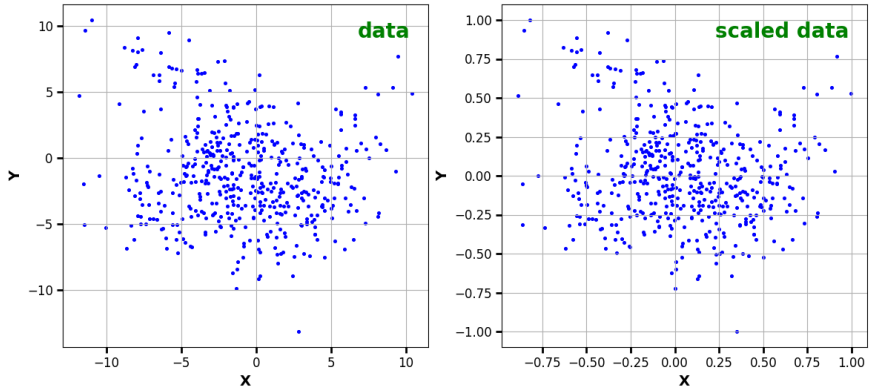


Figure 15: Data points sampled from one mog distribution with 10 components, MI between X and Y is 0.071.