
AFRL-RI-RS-TR-2019-186

# DATA-EFFICIENT NEURAL MUTUAL INFORMATION ESTIMATION FOR CAPTURING BRAIN-TO-BRAIN COMMUNICATION

SRI INTERNATIONAL

*SEPTEMBER 2019*

FINAL TECHNICAL REPORT

***APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED***

STINFO COPY

**AIR FORCE RESEARCH LABORATORY**
**INFORMATION DIRECTORATE**

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

# NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th ABW, Wright-Patterson AFB Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2019-186 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

**/ S /**
STEVEN DRAGER
Work Unit Manager

**/ S /**
DONALD TELESCA
Acting Technical Advisor, Computing
and Communication Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

# REPORT DOCUMENTATION PAGE

*Form Approved*
**OMB No. 0704-0188**



| 1. REPORT DATE *(DD-MM-YYYY)* | 2. REPORT TYPE | 3. DATES COVERED *(From - To)* |
|---|---|---|
| SEPTEMBER 2019 | FINAL TECHNICAL REPORT | OCT 2017 – APR 2019 |

**4. TITLE AND SUBTITLE**

DATA-EFFICIENT NEURAL MUTUAL INFORMATION ESTIMATION FOR CAPTURING BRAIN-TO-BRAIN COMMUNICATION

**5a. CONTRACT NUMBER**
FA8750-18-C-0213

**5b. GRANT NUMBER**
N/A

**5c. PROGRAM ELEMENT NUMBER**
61101E

**6. AUTHOR(S)**

Xiao Lin, Indranil Sur, Ajay Divakaran, Mohamed Amer, Sam Nastase, Uri Hasson

**5d. PROJECT NUMBER**
DSD2

**5e. TASK NUMBER**
SS

**5f. WORK UNIT NUMBER**
RI

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

**PRIME**
SRI International
201 Washington Rd.
Princeton NJ 08540

**SUB**
Princeton University
Department of Psychology and Neuroscience
Princeton, NJ 08540

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

Air Force Research Laboratory/RITA
525 Brooks Road
Rome NY 13441-4505

DARPA
675 North Randolph Street
Arlington, VA 22203-2114

**10. SPONSOR/MONITOR'S ACRONYM(S)**
AFRL/RI

**11. SPONSOR/MONITOR'S REPORT NUMBER**
AFRL-RI-RS-TR-2019-186

**12. DISTRIBUTION AVAILABILITY STATEMENT**

Approved for Public Release; Distribution Unlimited. PA# 88ABW-2019-4406
Date Cleared: 16 SEP 2019

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**

Measuring Mutual Information (MI) between high-dimensional, continuous, random variables from observed samples has wide theoretical and practical applications. Traditional MI methods, capable of capturing MI between low-dimensional signals, fall short when dimensionality increases and are not scalable. Existing neural approaches search for a d-dimensional neural network that maximizes a variational lower bound for mutual information estimation; however, this requires O(d log d) observed samples to prevent the neural network from overfitting. For practical mutual information estimation in real world applications, data is not always available at a surplus, especially in cases where acquisition of the data is prohibitively expensive, for example in fMRI analysis. This effort introduces a scalable, data-efficient mutual information estimator. BY coupling a learning-based view of the MI lower bound with meta-learning, NeuralMI achieves high-confidence estimations irrespective of network size and with improved accuracy at practical dataset sizes. The effectiveness has been demonstrated on synthetic benchmarks as well as a real world application of fMRI inter-subject correlation analysis.

**15. SUBJECT TERMS**
Mutual Information Estimation, Neural Networks, fMRI

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | **STEVEN DRAGER** |
| U | U | U | UU | 36 | 19b. TELEPHONE NUMBER (*Include area code*) **NA** |

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENTS

This work is funded by the Defense Advanced Research Projects Agency (DARPA) under Air Force Research Laboratory (AFRL) contract FA8750-18-C-0213. The views, opinions, and/or conclusions contained in this report are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied of the DARPA or the Department of Defense (DoD).

# 1.0 SUMMARY

Measuring Mutual Information (MI) between high-dimensional, continuous, random variables from observed samples has wide theoretical and practical applications. Recent work, Mutual Information Neural Estimation [1] (*MINE*), focused on estimating tight variational lower bounds of MI using neural networks, but assumed unlimited supply of samples to prevent overfitting. In real world applications, data is not always available at a surplus. In this work, we focus on improving data efficiency and propose a Data-Efficient MINE Estimator (DEMINE), by developing a relaxed predictive MI lower bound that can be estimated at higher data efficiency by orders of magnitudes. The predictive MI lower bound also enables us to develop a new meta-learning approach using task augmentation, Meta-learned Data-Efficient MINE Estimator (Meta-DEMINE) to improve generalization of the network and further boost estimation accuracy empirically. With improved data-efficiency, our estimators enable statistical testing of non-linear dependency at practical dataset sizes, which extends and supplements statistical testing of linear correlation commonly used in neuroscience and psychology for analyzing experiment results. We demonstrate the effectiveness of our estimators on synthetic benchmarks and a real world functional Magnetic Resonance Imaging (fMRI) dataset, with application of inter-subject correlation (ISC) analysis of measuring brain-to-brain coupling. Figure 1 provides an overview of our DEMINE approach.

➢ A statistical test of dependency is the non-linear extension of the widely used statistical test of correlation

Statistically significant dependency (mutual information)

$$I(X,Y) = \mathbb{E}_{XY} \log \frac{P(X,Y)}{P(X)P(Y)} > 0$$

* Here, $\rho(X,Y) = 0$

Statistically significant correlation

$$\rho(X,Y) = \mathbb{E}_{XY} \frac{(X-\bar{X})(Y-\bar{Y})}{\sigma_X \sigma_Y} \neq 0$$

➢ DEMINE enables statistical testing of dependency under practical sample sizes

| | Detecting non-linear interactions | Accurate signal modeling using deep nets | Statistical testing under practical sample size |
|---|---|---|---|
| Pearson's correlation | ✗ | ✗ | ✓ |
| KNN-based MI estimation (Kraskov et al. 2004) | ✓ | ✗ | ✗ |
| MINE-based MI estimation (Belghazi et al. 2018) | ✓ | ✓ | ✗ |
| **Our approach: Data Efficient MINE (DEMINE)** | ✓ | ✓ | ✓ |

➢ On synthetic benchmarks, DEMINE allows statistical testing of non-linear dependency under as few as 300 samples

| 300 samples of | 20-D Synthetic Gaussian data | 1-D Synthetic sine wave data |
|---|---|---|
| |  |  |
| (Confidence interval 95%) | | |
| Correlation | $\boldsymbol{\rho \approx 0.472 \pm 0.077}$* | $\rho \approx 0.069 \pm 0.104$* |
| Previous work (MINE-f) | $MI \geq 0.854 \pm 54.356$ | $MI \geq 0.768 \pm 65.843$ |
| **Our Approach (DEMINE)** | $\boldsymbol{MI \geq 0.038 \pm 0.030}$ | $\boldsymbol{MI \geq 0.382 \pm 0.187}$ |

* Correlation confidence interval is asymmetric. Reporting the end closer to 0

➢ On fMRI dataset, DEMINE/Meta-DEMINE identifies **5/6 additional brain regions** that have statistically significant non-linear dependency but lack statistically significant linear correlation. ($p<0.05$)

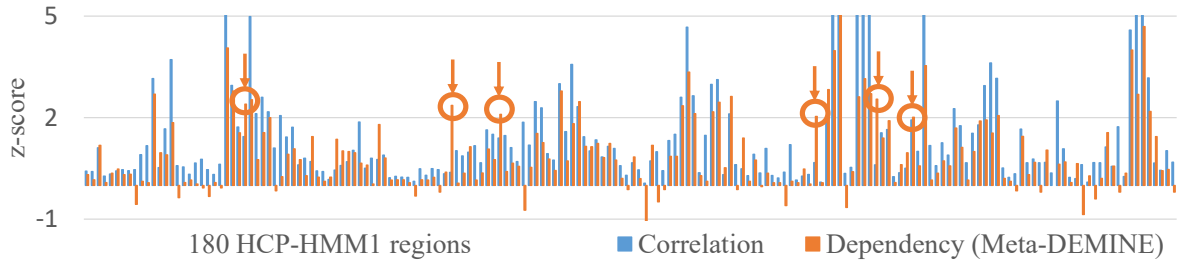



**Figure 1: An overview of our DEMINE approach, a data-efficient mutual information estimator that enables statistical test of dependency at practical sample sizes. DEMINE is able to find statistically significant dependency using as few as 300 samples on synthetic benchmarks, and complements statistical test of correlation on a real fMRI dataset of brain-to-brain coupling.**

# 2.0 INTRODUCTION

Mutual Information is an important, theoretically grounded, measure of similarity between random variables. MI captures general, non-linear, statistical dependencies between random variables. It is a widely used quantity in various machine learning tasks ranging from classification to feature selection and neural network analysis.

A widely used approach for estimating MI from samples is using k-nearest neighbors (k-NN) estimates, notably the Kraskov-Stogbauer-Grassberger (KSG) estimator [2]. Recent work [3] provided a comprehensive review and studied the consistency and asymptotic confidence bound of the KSG estimator [4]. MI estimation can also be achieved by estimating individual entropy terms involved through kernel density estimation [5] or cross-entropy [6]. Overfitting can be reduced through partitioning the samples into different folds for modeling and for estimation. Despite their fast and accurate estimations on random variables with few dimensions, MI estimation on high-dimensional random variables remains challenging for commonly used Gaussian kernels. Fundamentally, estimating MI requires the ability to accurately model the random variables, where high-capacity neural networks have shown excellent performance on complex high-dimensional signals such as text, image and audio.

Recent works on MI estimation have focused on developing tight variational MI lower bounds where neural networks are used for signal modeling. The Information Maximization (IM) algorithm [7] introduces a variational MI lower bound, where a neural network $q(z|x)$ is learned as a variational approximation to the conditional distribution $P(Z|X)$. Here $X$ and $Z$ are random variables whose mutual information – denoted as $I(X;Z)$ – is the objective we want to compute, and $x$ and $z$ denote samples of $X$ and $Z$. The IM algorithm requires that the entropy, $H(Z)$, as well as the quantity $E_{XZ} \log q(z|x)$ to be tractable. This is true for latent codes of Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) as well as categorical variables. Recent work [1] introduces MI lower bounds *MINE* and *MINE-f* that allow the modeling of general random variables without restrictions and shows improved accuracy on high-dimensional random variables, with application to improving generative models. [8] introduces a spectrum of energy-based MI estimators based on *MINE* and *MINE-f* lower bounds, as well as a new estimator named TCPC for the case where it's possible to draw multiple samples from $P(Z|X)$.

An important challenge that previous works overlooked is MI estimation using limited data. Because the high-capacity neural networks tend to overfit, variational estimators, such as *MINE*, expect an impractically large number of samples to overcome overfitting and to reach high confidence. In addition, tighter lower bounds may also require more data to estimate. When merely a few samples are provided, it was observed that the resulting estimations suffer from high variance observed in [8].

To address the data efficiency challenge, our estimator, DEMINE, introduces predictive mode and meta-learning to the *MINE* estimator family to greatly improve sample efficiency. We develop a relaxed, predictive variational lower bound based on *MINE* that prevents overfitting by explicitly partitioning samples into training and validation. Furthermore, our predictive formulation allows us to incorporate techniques that improves generalization beyond curve fitting such as meta-learning. With these improvements, we show that DEMINE enables practical statistical testing of dependency in not only synthetic datasets but also for real world fMRI data analysis for capturing nonlinear and

higher-order brain-to-brain coupling.

An additional component to enhance our estimators is meta-learning. Meta-learning, or "learning to learn", seeks to improve the generalization capability of neural networks by searching for better hyper parameters [9], network architectures [10], initialization [11, 12, 13] and distance metrics [14, 15]. Meta-learning approaches have shown significant performance improvements in applications such as automatic neural architecture search [10], few-shot image recognition [11] and imitation learning [16].

In particular, our estimator benefits from the Model-Agnostic Meta-Learning (MAML) [11] framework which is designed to improve few-shot learning performance. In few-shot learning, the task is learning to classify an input image into K categories. For each category only N images are provided for training as the "support set". But additional training images may be provided for other categories as the "training set". The challenge of few-shot learning is how to effectively leverage training set such that new classifiers learned on the support set can achieve high performance. MAML learns a network initialization by maximizing the performance after fine-tuning the network from the initialization on the support set. Applications of MAML include few-shot image classification and navigation. In this work, we propose an approach to use MAML for the new task of maximizing MI lower bounds. The model-agnostic nature of MAML allows our method to be applied to generic random variables. To construct a collection of diverse tasks for MAML learning from limited samples, inspired by MI's invariance to invertible transformations, we propose a task-augmentation protocol to automatically construct MI estimation tasks by sampling random transformations to transform the samples. Results show reduced overfitting and improved generalization.

Our contributions are summarized as follows: 1) Data Efficient Mutual Information Neural Estimator enabling statistical test of dependency; 2) New formulation of meta-learning using Task Augmentation (Meta-DEMINE); 3) Application to real life, data scarse application (fMRI brain-to-brain coupling analysis).

# 3.0 METHODS, ASSUMPTIONS AND PROCEDURES

## 3.1 Mutual Information Estimation: Background

In this section, we will provide the background necessary to understand our approach[1]. We define $X$ and $Z$ to be two random variables, $P(X, Z)$ is the joint distribution, and $P(X)$ and $P(Z)$ are the marginal distributions over $X$ and $Z$ respectively. Our goal is to estimate MI, $I(X; Z)$ given independent and identically distributed (*i.i.d.*) sample pairs $(x_i, z_i)$, $i = 1, 2 \dots n$ from $P(X, Z)$. Let $\mathcal{F} = \{T_\theta(x, z)\}_{\theta \in \Theta}$ be a class of scalar functions, where $\theta$ is the set of model parameters. Let $q(x|z) = p(x)\frac{e^{T_\theta(x,z)}}{\mathbb{E}_{(x,z)\sim P_{XZ}} e^{T_\theta(x,z)}}$. Results from previous works [1, 8] show that the following energy-based family of lower bounds of MI hold for any $\theta$:

$$\begin{aligned} I(X; Z) &\geq \mathbb{E}_{(x,z)\sim P_{XZ}} \log \tfrac{q(x|z)}{p(x)} = \mathbb{E}_{(x,z)\sim P_{XZ}} T_\theta(x, z) - \mathbb{E}_{x\sim P_X} \log \mathbb{E}_{z\sim P_Z} e^{T_\theta(x,z)} \triangleq I_{\text{EB1}} \text{ [8]} \\ &\geq \mathbb{E}_{(x,z)\sim P_{XZ}} T_\theta(x, z) - \log \mathbb{E}_{x\sim P_X, z\sim P_Z} e^{T_\theta(x,z)} \triangleq I_{\text{MINE}} \text{ [1]} \\ &\geq \mathbb{E}_{(x,z)\sim P_{XZ}} T_\theta(x, z) - \mathbb{E}_{x\sim P_X, z\sim P_Z} e^{T_\theta(x,z)} + 1 \triangleq I_{\text{MINE-f}} \text{ [1]}, I_{\text{EB}} \text{ [8]} \end{aligned} \tag{1}$$

where, $\mathbb{E}$ is the expectation over the given distribution. Based on $I_{\text{MINE}}$, the *MINE* estimator $\widehat{I(X, Z)}_n$ is defined as in Eq.2. Estimators for $I_{\text{EB1}}$, $I_{\text{MINE-f}}$ and $I_{\text{EB}}$ can be defined similarly.

$$\widehat{I(X; Z)}_n = \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} T_\theta(x_i, z_i) - \log \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} e^{T_\theta(x_i, z_j)}. \tag{2}$$

With infinite samples to approximate expectation, Eq.2 converges to the lower bound $\widehat{I(X, Z)}_\infty = \sup_{\theta \in \Theta} I_{\text{MINE}}$. Note that the number of samples $n$ needs to be substantially more than the number of model parameters $d = |\theta|$ to prevent $T_\theta(X, Y)$ from overfitting to the samples $(x_i, z_i)$, $i = 1, 2 \dots n$ and overestimating MI. Formally, the sample complexity of *MINE* is defined as the minimum number of samples $n$ in order to achieve Eq.3,

$$\Pr(|\widehat{I(X, Z)}_n - \widehat{I(X, Z)}_\infty| \leq \epsilon) \geq 1 - \delta. \tag{3}$$

Specifically, *MINE* proves that under the following assumptions: 1) $T_\theta(X, Z)$ is $L$-Lipschitz; 2) $T_\theta(X, Z) \in [-M, M]$, 3) $\{\theta_i \in [-K, K], \quad \forall i \in 1, \dots, d\}$, the sample complexity of *MINE* is given by Eq.4.

$$n \geq \frac{2M^2(d \log(16KL\sqrt{d}/\epsilon) + 2dM + \log(2/\delta))}{\epsilon^2}. \tag{4}$$

For example, a neural network with dimension $d = 10,000$, $M = 1$, $K = 0.1$ and $L = 1$, achieving a confidence interval of $\epsilon = 0.1$ with $95\%$ confidence ($\delta = 0.05$) would require $n \geq 18,756,256$ samples. This is achievable for synthetic example generated by GANs like that studied in [1]. For real data, however, the cost of data acquisition for reaching statistically significant estimation can be prohibitively expensive. We propose to use the MI lower bounds specified in Eq.1 from a prediction perspective, inspired by cross-validation. Our estimator, DEMINE, improves sample complexity by disentangling data for lower bound estimation from data for learning a generalizable $T_\theta(X, Z)$. DEMINE enables high-confidence MI estimation on small datasets. Bound tightness is further improved by Meta-DEMINE by using meta-learning to learn generalizable $T_\theta(X, Z)$.

[1]We follow the same notation in [1]. We encourage the review of [1, 8] for a detailed understanding of $I_{\text{MINE}}$, $I_{\text{EB1}}$, and $I_{\text{EB}}$.

## 3.2 Approach

§3.2.1 specifies DEMINE for predictive MI estimation and derives the confidence interval; §3.2.2 formulates Meta-DEMINE, explains task augmentation, and defines the optimization algorithms.

### 3.2.1 Predictive Mutual Information Estimation

In DEMINE, we interpret the estimation of *MINE-f* lower bound[2] Eq.1 as a learning problem. The goal is to infer the optimal network $T_{\theta^*}(X, Z)$ with parameters $\theta^*$ using a limited number of samples defined as follows:

$$\theta^* = \arg\max_{\theta\in\Theta} \mathbb{E}_{P_{XZ}} T_\theta(X, Z) - \mathbb{E}_{P_X}\mathbb{E}_{P_Z} e^{T_\theta(X,Z)} + 1.$$

Specifically, samples from $P(X, Z)$ are subdivided into a training set $\{(x_i, z_i)_{\text{train}}, i = 1, \ldots, m\}$ and a validation set $\{(x_i, z_i)_{\text{val}}, i = 1, \ldots, n\}$. The training set is used for learning a network $\tilde{\theta}$ as an approximation to $\theta^*$ whereas the validation set is used for computing the DEMINE estimation $\widehat{I(X, Z)}_{n,\tilde{\theta}}$ defined as in Eq.5.

$$\widehat{I(X, Z)}_{n,\tilde{\theta}} = \frac{1}{n}\sum_{i=1}^{n} T_{\tilde{\theta}}(x_i, z_i)_{\text{val}} - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{T_{\tilde{\theta}}(x_i,z_j)_{\text{val}}} + 1 \tag{5}$$

We propose an approach to learn $\tilde{\theta}$, DEMINE. DEMINE learns $\tilde{\theta}$ by maximizing the MI lower bound on the training set as follows:

$$\tilde{\theta} = \arg\min_{\theta\in\Theta} \mathcal{L}(\{(x, z)\}_{\text{train}}, \theta), \text{ where,}$$

$$\mathcal{L}(\{(x, z)\}_{\mathcal{B}}, \theta) = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|} T_\theta(x_i, z_i)_{\mathcal{B}} + \frac{1}{|\mathcal{B}|^2}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|} e^{T_\theta(x_i,z_j)_{\mathcal{B}}} - 1. \tag{6}$$

The DEMINE algorithm is shown in Algorithm 1.

**Sample complexity analysis.** Because $\tilde{\theta}$ is learned independently of validation samples $\{(x_i, z_i)_{\text{val}}, i = 1, \ldots, n\}$, the sample complexity of the DEMINE estimator does not involve the model class $\mathcal{F}$ and the sample complexity is greatly reduced compared to *MINE-f*. DEMINE estimates $\widehat{I(X, Z)}_{\infty,\tilde{\theta}}$ when infinite number of samples are provided, defined as:

$$\begin{aligned} \widehat{I(X, Z)}_{\infty,\tilde{\theta}} &= \mathbb{E}_{P_{XZ}} T_{\tilde{\theta}}(X, Z) - \mathbb{E}_{P_X}\mathbb{E}_{P_Z} e^{T_{\tilde{\theta}}(X,Z)} + 1 \\ &\leq \sup_{\theta\in\Theta} \mathbb{E}_{P_{XZ}} T_\theta(X, Z) - \mathbb{E}_{P_X}\mathbb{E}_{P_Z} e^{T_\theta(X,Z)} + 1 \leq I(X; Z) \end{aligned} \tag{7}$$

We now derive the sample complexity of DEMINE defined as the number of samples $n$ required for $\widehat{I(X, Z)}_{n,\tilde{\theta}}$ to be a good approximation to $\widehat{I(X, Z)}_{\infty,\tilde{\theta}}$ in Theorem 1.

[2] *MINE* lower bound can also be interpreted in the predictive way, but will result in a higher sample complexity than *MINE-f* lower bound. We choose *MINE-f* in favor of a lower sample complexity over bound tightness.

**Algorithm 1** DEMINE

**Input Data**: $\{(x,z)_{\text{train}}, (x,z)_{\text{val}}\}$
**Parameters**: Batch $\mathcal{B}$, Iterations $N_O$, Learning rate $\eta$
**Output**: MI, $T_\theta(X,Z)$

$\theta^{(0)} \leftarrow$ Xavier Initialization [17]
**for** i = 1 : $N_O$ **do**
    Sample a batch of $(x_i, z_i)_{\mathcal{B}} \sim (x,z)_{\text{train}}$
    Compute $\mathcal{L}\left((x_i, z_i)_{\mathcal{B}}, \theta^{(i-1)}\right)$
    Compute $\nabla_\theta^{(i)}\mathcal{L}$ – gradient for $\theta$
    Update $\theta^{(i)}$ using Adam [20] with $\eta$
**end for**
MI $= \widehat{I(X,Z)}_{n,\theta^{(N_O)}}$
**return** MI, $\theta^{(N_O)}$

**Theorem 1.** For $T_{\tilde{\theta}}(X,Z)$ bounded by $[L,U]$, given any accuracy $\epsilon$ and confidence $\delta$, we have:

$$\Pr(|\widehat{I(X,Z)}_{n,\tilde{\theta}} - \widehat{I(X,Z)}_{\infty,\tilde{\theta}}| \leq \epsilon) \geq 1-\delta$$

when the number of validation samples $n$ satisfies:

$$n \geq n^*, s.t.\ f(n^*) \equiv \min_{0\leq\xi\leq\epsilon} 2e^{-\frac{2\xi^2 n^*}{(U-L)^2}} + 4e^{-\frac{(\epsilon-\xi)^2 n^*}{2(e^U-e^L)^2}} = \delta \tag{8}$$

**Proof.** Since $T_{\tilde{\theta}}(X,Z)$ is bounded by$[L,U]$, applying the Hoeffding inequality [18] to the first half of Eq.5 yields:

$$\Pr(|\frac{1}{n}\sum_{i=1}^{n} T_{\tilde{\theta}}(x_i,z_i) - \mathbb{E}_{P_{XZ}} T_{\tilde{\theta}}(X,Z)| \geq \xi) \leq 2e^{-\frac{2\xi^2 n}{(U-L)^2}}$$

As $e^{T_\theta(X,Z)}$ is bounded by $[e^L, e^U]$, applying the Hoeffding inequality twice to the second half of Eq.5:

$$\begin{aligned}
\Pr(|\mathbb{E}_{P_X}\mathbb{E}_{P_Z} e^{T_\theta(X,Z)} - \tfrac{1}{n}\textstyle\sum_{i=1}^n \mathbb{E}_{P_Z} e^{T_{\tilde{\theta}}(x_i,z)}| \geq \zeta) &\leq 2e^{-\frac{2\zeta^2 n}{(e^U-e^L)^2}} \\
\Pr(|\mathbb{E}_{P_Z}\tfrac{1}{n}\textstyle\sum_{i=1}^n e^{T_\theta(x_i,z)} - \tfrac{1}{n}\textstyle\sum_{j=1}^n \tfrac{1}{n}\textstyle\sum_{i=1}^n e^{T_{\tilde{\theta}}(x_i,z_j)}| \geq \zeta) &\leq 2e^{-\frac{2\zeta^2 n}{(e^U-e^L)^2}}
\end{aligned}$$

Combining the above bounds results in:

$$\Pr(|\widehat{I(X,Z)}_{n,\tilde{\theta}} - \widehat{I(X,Z)}_{\infty,\tilde{\theta}}| \leq \xi + 2\zeta) \geq 1 - 2e^{-\frac{2\xi^2 n}{(U-L)^2}} - 4e^{-\frac{2\zeta^2 n}{(e^U-e^L)^2}}$$

By solving $\xi$ to minimize $n$ according to Eq.8 we have:

$$\Pr(|\widehat{I(X,Z)}_{n,\tilde{\theta}} - \widehat{I(X,Z)}_{\infty,\tilde{\theta}}| \leq \epsilon) \geq 1-\delta. \qquad \blacksquare$$

Compared to *MINE*, as per the example shown in §3.1, for $M=1$ (*i.e.* $L=-1$ and $U=1$), $\delta = 0.05$, $\epsilon = 0.1$, our estimator requires $n = 10,742$ compared to *MINE* requiring $n = 18,756,256$ *i.i.d* validation samples to estimate a lower bound, which makes MI-based dependency analysis

feasible for domains where data collection is prohibitively expensive, *e.g.* fMRI brain scans. In practice, sample complexity can be further optimized by tuning hyperparameters $U$ and $L$.

Note that the sample complexity of our approach, DEMINE, for estimating Eq.7 does not depend on network size $d$. The improved sample complexity seemingly comes at a cost of bound tightness guarantees. In fact, to guarantee bound tightness of Eq.7, $O(d \log d)$ examples would still be theoretically required to learn $\tilde{\theta}$ with guaranteed close values to $\theta^*$, and the total data cost would be on par with *MINE*. In practice, such a learnability bound is known to be overly loose, as over-parameterized neural networks have been shown to generalize well in classification and regression tasks [19]. Fundamentally, what determines bound tightness is the generalization error of $\tilde{\theta}$ – to which the learnability bound is serving as a proxy. Empirically, not only that the bound tightness of DEMINE is as good as *MINE* so the loss of guaranteed tightness did not affect empirical tightness, but the learning-based formulation of DEMINE also allows further bound tightness improvements by learning $\tilde{\theta}$ that generalizes beyond curve fitting using meta-learning.

In the following section, we present a meta-learning formulation, Meta-DEMINE, that learns $\tilde{\theta}$ for generalization given the same model class and training samples.

### 3.2.2 Meta-Learning

Given training data $\{(x_i, z_i)_{\text{train}}, i = 1, \dots m\}$, Meta-DEMINE algorithm first generates MI estimation tasks each consisting of a meta-training split A and a meta-val split B through a novel *task augmentation* process. A parameter initialization $\theta_{\text{init}}$ is then learned to maximize MI estimation performance on the generated tasks using initialization $\theta_{\text{init}}$ as shown in Eq.9.

$$\theta_{\text{init}} = \arg\min_{\theta^{(0)} \in \Theta} \mathbb{E}_{(\text{A,B}) \in \mathcal{T}} \mathcal{L}((x, z)_{\text{B}}, \theta^{(t)}), \text{with }, \theta^{(t)} \equiv \text{MetaTrain}\big((x, z)_{\text{A}}, \theta^{(0)}\big). \tag{9}$$

Here $\theta^{(t)} = \text{MetaTrain}\big((x, z)_{\text{A}}, \theta^{(0)}\big)$ is the meta-training process of starting from an initialization $\theta^{(0)}$ and applying Stochastic Gradient Descent (SGD) [3] over $t$ steps to learn $\theta$ where in every meta training iteration we have:

$$\theta^{(t)} \leftarrow \theta^{(t-1)} - \gamma \nabla \mathcal{L}((x, z)_{\text{A}}, \theta^{(t-1)}).$$

Finally, $\tilde{\theta}$ is learned using the entire training set $\{(x_i, z_i)_{\text{train}}, i = 1, \dots, m\}$ with $\theta_{\text{init}}$ as initialization:

$$\tilde{\theta} = \text{MetaTrain}\big((x, z)_{\text{train}}, \theta_{\text{init}}\big).$$

**Task Augmentation:** Meta-DEMINE adapts MAML [11] for MI lower bound maximization. MAML has been shown to improve generalization performance in $N$-class $K$-shot image classification. MI estimation, however, does not come with predefined classes and tasks. A naive approach to produce tasks would be through cross validation – partitioning training data into meta-training and meta-validation splits. However, merely using cross-validation tasks is prone to overfitting – a $\theta_{\text{init}}$, which memorizes all training samples would as a result have memorized all meta-validation splits. Instead, Meta-DEMINE generates tasks by augmenting the cross validation tasks through

[3] In practice, the Adam optimizer [20] is used for faster optimization. The Adam optimizer uses first and second order momentum of the gradient to speed up optimization. Illustrating SGD for simplicity.

*task augmentation*. Training samples are first split into meta-training and meta-validation splits, and then transformed using the same random invertible transformation to increase task diversity. Meta-DEMINE generates invertible transformation by sequentially composing the following functions:

$$
\begin{array}{lll}
\textit{Mirror}: & m(x) = (2n-1)x, & n \sim \text{Bernoulli}(\frac{1}{2}), \\
\textit{Permute}: & P(x) = {}^{n}P_{d}, & \text{Permute dimensions.} \\
\textit{Offset}: & O(x) = x + \epsilon, & \epsilon \sim \mathcal{U}(-0.1, 0.1), \\
\textit{Gamma}: & G(x) = \text{sign}(x)\,|x|^{\gamma}, & \gamma \sim \mathcal{U}(0.5, 2),
\end{array}
$$

Since the MI between two random variables is invariant to invertible transformations on each variable, MetaTrain is expected to arrive at the same MI lower bound estimation regardless of the transformation applied. At the same time, memorization is greatly suppressed, as the same pair $(x, z)$ can have different $\log \frac{p(x,z)}{p(x)p(z)}$ under different transformations. More sophisticated invertible transformations (affine, piece-wise linear) can also be added. Task augmentation is an orthogonal approach to data augmentation. Using image classification as an example, data augmentation generates variations of the image, translated, or rotated images assuming that they are valid examples of the class. Task augmentation on the other hand, does not make such an assumption. Task augmentation requires the initial parameters $\theta_{\text{init}}$ to be capable of recognizing the same class in a world where all images are translated and/or rotated, with the assumption that the optimal initialization should easily adapt to both the upright world and the translated and/or rotated world.

**Optimization:** Solving $\theta_{\text{init}}$ using the meta-learning formulation Eq.9 poses a challenging optimization problem. The commonly used approach is back propagation through time (BPTT) which computes second order gradients and directly back propagate gradient from MetaTrain$((x, z)_{\text{A}}, \theta^{(0)})$ to $\theta_{\text{init}}$. BPTT is very effective for a small number of optimization steps, but is vulnerable to exploding gradients and is memory intensive. In addition to BPTT, we find that stochastic finite difference algorithms such as Evolution Strategies (ES) [21] and Parameter-Exploring Policy Gradients (PEPG) [22] can sometimes improve optimization robustness. In practice, we use BPTT or PEPG to optimize Eq.9 depending on the problem. Meta-DEMINE algorithm is specified in Algorithm 2.

**Algorithm 2** Meta-DEMINE

**Input Data**: $\{(x, z)_{\text{train}}, (x, z)_{\text{val}}\}$
**Parameters**: batch $\mathcal{B}$, Meta Learning Iterations $N_M$, Task Augmentation Iterations $N_T$, Optimization Iterations $N_O$, Ratio $r$, Learning rate $\eta$, Meta Learning Rate $\eta_{\text{meta}}$
**Output**: MI, $T_{\theta_{\text{init}}}(X, Z)$, $T_\theta(X, Z)$

**for** i = 1 : $N_M$ **do**
  **for** j = 1 : $N_T$ **do**
    $A = r \times \text{train}, B = \text{train} - A$
    Split $(x, z)_{\text{train}}$ into $(x, z)_{\text{A}}$ and $(x, z)_{\text{B}}$
    Transformation $R_x$ for $x$, $R_x(\cdot) = \text{m}(\text{P}(\text{O}(\text{G}(\cdot))))$
    Transformation $R_z$ for $z$, $R_z(\cdot) = \text{m}(\text{P}(\text{O}(\text{G}(\cdot))))$
    $\theta^{(0)}_{\text{meta}} \leftarrow \theta_{\text{init}}$
    **for** k = 1 : $N_O$ **do**
      Sample a batch of $(x, z)_{\mathcal{B}} \sim (x, z)_{\text{A}}$
      Compute $\mathcal{L}\big((R_x(x), R_z(z))_{\mathcal{B}}, \theta^{(k)}_{\text{meta}}\big)$
      Compute $\nabla_{\theta^{(k)}_{\text{meta}}} \mathcal{L}$ – gradient for $\theta_{\text{meta}}$
      Update $\theta_{\text{meta}}$ using Adam [20] with $\eta$
    **end for**
    Compute $\mathcal{L}_{\text{meta}}\big((R_x(x), R_z(z))_{\text{B}}, \theta^{(N_O)}_{\text{meta}}\big)$
    Compute $\nabla_{\theta_0} \mathcal{L}_{\text{meta}}$ – gradient to $\theta_{\text{init}}$ using BPTT
  **end for**
  Update $\theta_{\text{init}}$ using Adam [20] with $\eta_{meta}$
**end for**
$\theta^{(0)} \leftarrow \theta_{\text{init}}$
**for** i = 1 : $N_O$ **do**
  Sample a batch of $(x, z)_{\mathcal{B}} \sim (x, z)_{\text{train}}$
  Compute $\mathcal{L}\big((x, z)_{\mathcal{B}}, \theta^{(i)}\big)$
  Compute gradient $\nabla_\theta \mathcal{L}$
  Update $\theta$ using Adam with $\eta$
**end for**
Compute MI $= \mathcal{L}\big((x, z)_{\text{val}}, \theta^{(N_O)}\big)$
**return** MI, $\theta_{\text{init}}$, $\theta^{(N_O)}$

# 4.0 RESULTS AND DISCUSSION

## 4.1 Evaluation on Synthetic Datasets

**Dataset.** We evaluate our approaches DEMINE and Meta-DEMINE against baselines and state-of-the-art approaches on 3 synthetic datasets: 1-dimensional (1D) Gaussian, 20-dimensional (20D) Gaussian and sine wave. For 1D and 20D Gaussian datasets, following [1], we define two $k$-dimensional multivariate Gaussian random variables $X$ and $Z$ which have component-wise correlation $corr(X_i, Z_j) = \delta_{ij}\rho$, where $\rho \in (-1, 1)$ and $\delta_{ij}$ is Kronecker's delta. Mutual information $I(X; Z)$ has a closed form solution $I(X; Z) = -k \ln(1 - \rho^2)$. For sine wave dataset, we define two random variables $X$ and $Z$, where $X \sim \mathcal{U}(-1, 1)$, $Z = \sin(aX + \frac{\pi}{2}) + 0.05\epsilon$, and $\epsilon \sim \mathcal{N}(0, 1)$. Estimating mutual information accurately given few pairs of $(X, Z)$ requires the ability to extrapolate the sine wave given few examples. Ground truth MI for sine wave dataset is approximated by running the the KSG Estimator [2] on $1,000,000$ samples.

**Implementation.** We compare our estimators, DEMINE and Meta-DEMINE, against the KSG estimator [2] MI-KSG and MINE-f. For both DEMINE and Meta-DEMINE, we study variance reduction mode, referred to as *-vr*, where hyperparameters are selected by optimizing 95% confident estimation mean ($\mu - 2\sigma_\mu$) and statistical significance mode, referred to as *-sig*, where hyperparameters are selected by optimizing 95% confident MI lower bound ($\mu - \epsilon$). Samples $(x, z)$ are split into 50%-50% as $(x, z)_{\text{train}}$ and $(x, z)_{\text{val}}$.

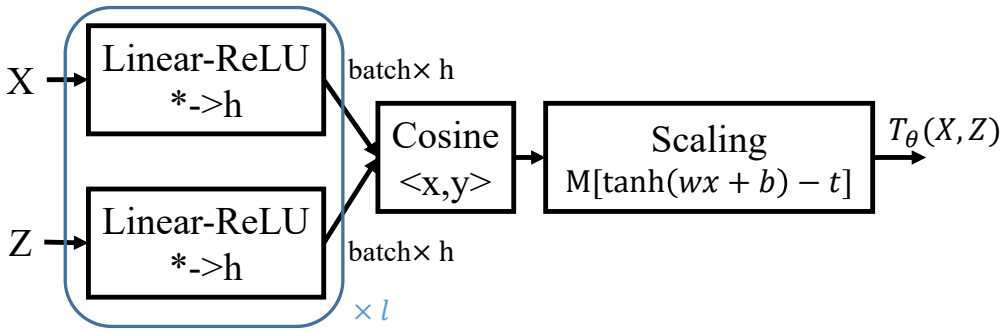



**Figure 2: Network architecture used for synthetic experiments. Incoming samples of random variables $X$ and $Z$ are encoded using MLP and are combined using cosine distance followed by a scaling layer.**

We use a separable network architecture $T_\theta(x, z) = M\big(\tanh(w \cos\langle f(x), g(z)\rangle + b) - t\big)$. The network architecture is illustrated in Figure 2. $f$ and $g$ are Multi-layer perceptrons (MLPs) encoders using Rectified linear units (ReLU). They embed signals $x$ and $z$ into vector embeddings. Hyperparameters $t \in [-1, 1]$ and $M$ control upper and lower bounds $T_\theta(x, z) \in [-M(1+t), M(1-t)]$. Parameters $w$ and $b$ are learnable parameters. MLP design and optimization hyperparameters are selected using Bayesian hyperparameter optimization [23] with 3-fold cross-validation on $(x, z)_{\text{train}}$ over 1,000 iterations.

Hyperparameter search on DEMINE-vr and DEMINE-sig was conducted using the hyperopt package [4]. Seven hyper parameters were involved in hyperparameter search: 1) number of encoder layers $[1, 5]$, 2) encoder hidden size $[8, 256]$, 3) learning rate $\eta$ $[10^{-4}, 3 \times 10^{-1}]$ in log scale, 4) number of optimization iterations $N_O$ $[5, 200]$ (sine wave $[5, 5000]$) in log scale, 5) batch size $\mathcal{B}$ $[256, 1024]$, 6) $M$, $[10^{-3}, 5]$ in log scale, 7) $t$, $[-1, 1]$. Mean $\mu$ and sample standard deviation $\sigma$ of MI estiamte computed over 3 fold cross validation on $(x, z)_{\text{train}}$. DEMINE-vr maximizes two sigma low $\mu - 2\sigma_\mu$ where $\sigma_\mu = \frac{1}{\sqrt{3}}\sigma$ due to 3-fold crossval. DEMINE-sig maximizes statistical significance $\mu - \epsilon$ where $\epsilon$ is two-sided 95% confidence interval of MI. Meta-DEMINE-vr and Meta-DEMINE-sig subsequently reuse these hyperparameters as DEMINE-vr and DEMINE-sig.

Meta-learning hyperparameters are empirically chosen as outer loop $N_M = 3,000$ iterations, task augmentation $N_T = 1$ iterations, $r = 0.8$, $\eta_{\text{meta}} = \frac{\eta}{3}$, with task augmentation mode $m(P(O(\cdot)))$. $N_O$ capped at 30 iterations for 1D and 20D Gaussian datasets due to memory limit. The sine wave datasets seem to require large $N_O$. We use PEPG [22] rather than BPTT.

For MI-KSG, we use off-the-shelf implementation [3] with default number of nearest neighbors k=3. MI-KSG does not provide any confidence interval. For MINE-f, we use the same network architecture as DEMINE-vr. We implement both the original formulation which optimizes $T_\theta$ on $(x,\ z)$ till convergence (10,000 iterations), as well as our own implementation of MINE-f augmented with early stopping (MINE-f-ES), where optimization is stopped after the same number of iterations as DEMINE-vr to control overfitting.

**Results.** Figure 3(a) shows MI estimation performance on 20D Gaussian datasets with varying $\rho \in \{0,\ 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5\}$ using $N = 300$ samples. Results are averaged over 5 runs to compare estimator bias, variance and confidence. Estimator mean and 1-sigma estimator standard deviation are visualized as boxes, and $95\%$ confidence intervals are shown in upper and lower bound lines. Ground truth mutual information is shown in horizontal lines, and estimator bias is the difference between ground truth and estimator mean. From Figure 3(a), we can see that Meta-DEMINE-sig detects the highest $p < 0.05$ confidence MI, outperforming DEMINE-sig which is a close second. Both detect $p < 0.05$ statistically significant dependency starting $\rho = 0.3$, whereas estimations of all other approaches are low confidence. It shows that in contrary to common belief, estimating the variational lower bounds with high confidence can be challenging under limited data. When there is no correlation, *i.e.* $\rho = 0$, MINE-f gives an incorrect MI estimate MI$> 3.0$ and MINE-f-ES estimates positive MI, both due to overfitting, despite MINE-f-ES having the lowest empirical bias. DEMINE variants have relatively high empirical bias but low variance due to tight upper and lower bound control, which provides a different angle to understand bias-variance trade off in MI estimation [8].

Figure 3(b,c,d) shows MI estimation performance on 1D, 20D Gaussian and sine wave datasets with fixed $\rho = 0.8,\ 0.3$ and $a = 8\pi$ respectively, with varying $N \in \{30,\ 100,\ 300,\ 1000,\ 3000\}$ number of samples. More samples asymptotically improves empirical bias across all estimators. As opposed to 1D Gaussian datasets which are well solved by $N = 300$ samples, higher-dimensional 20D Gaussian and higher-complexity sine wave datasets are much more challenging and are not solved using $N = 3000$ samples with a signal-agnostic MLP architecture. DEMINE-sig and Meta-DEMINE-sig detect $p < 0.05$ statistically significant dependency on not

[4] Hyperopt package: `https://github.com/hyperopt/hyperopt`.

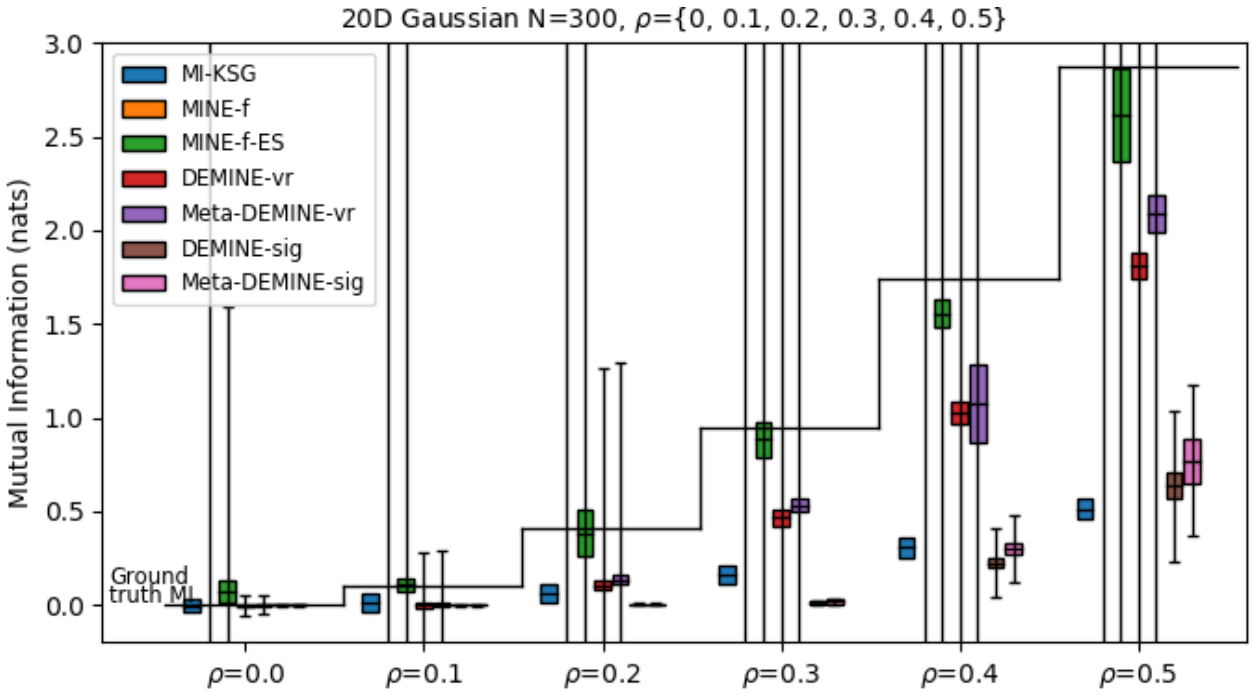
only 1D and 20D Gaussian datasets where $x$ and $z$ have non-zero correlation, but also on the sine wave datasets where correlation between $x$ and $z$ is 0. This means that DEMINE-sig and Meta-DEMINE-sig can be used for nonlinear dependency testing to complement linear correlation testing.

We study the effect of cross-validation meta-learning and task augmentation on 20D Gaussian with $\rho = 0.3$ and $N = 300$. Figure 4 plots performance of Meta-DEMINE-vr over $N_M = 3000$ meta iterations under combinations of task augmentations modes and number of adaptation iterations $N_O \in \{0, 20\}$. Overall, task augmentation modes which involve axis flipping $m(\cdot)$ and permutation $P(\cdot)$ are the most successful. With $N_O = 20$ steps of adaptation, task augmentation modes $P(\cdot)$, $m(P(\cdot))$ and $m(P(O(\cdot)))$ prevent overfitting and improves performance. The performance improvements of task augmentation is not simply from change in batch size, learning rate or number of optimization iterations, because meta-learning without task augmentation for both $N_O = 0$ and 20 could not outperform baseline. Meta-learning without task augmentation and with task augmentation but using only $O(\cdot)$ or $G(\cdot)$ result in overfitting. Task augmentation with $m(\cdot)$ or $m(P(O(G(\cdot))))$ prevent overfitting, but do not provide performance benefits, possibly because their complexity is insufficient or excessive for 20 adaptation steps. Further more, task augmentation with no adaptation ($N_O = 0$) falls back to data augmentation, where samples from transformed distributions are directly used to learn $T_\theta(x, z)$. Data augmentation with $O(\cdot)$ outperforms no augmentation, but is unable to outperform baseline and suffer from overfitting. It shows that task augmentation provides improvements orthogonal to data augmentation.

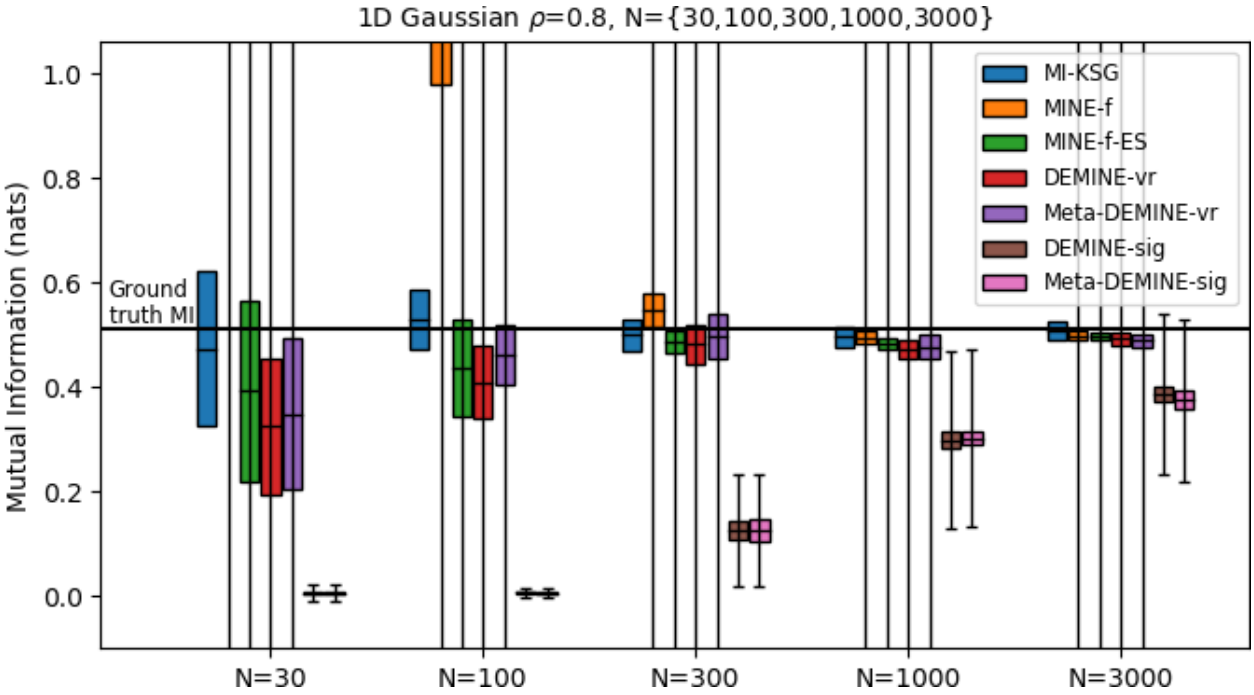
## 4.2 Application: fMRI Inter-Subject Correlation (ISC) Analysis

Humans use language to effectively transmit brain representations among conspecifics. For example, after witnessing an event in the world, a speaker may use verbal communication to evoke neural representations reflecting that event in a listener's brain [24]. The efficacy of this transmission, in terms of listener comprehension, is predicted by speaker–listener neural synchrony and synchrony among listeners [25]. To date, most work has measured brain-to-brain synchrony by locating statistically significant ISC; quantified as the Pearson product-moment correlation coefficient between response time series for corresponding voxels or regions of interest (ROIs) across individuals [26, 27, 28]. Using DEMINE and Meta-DEMINE for statistical dependency testing, we can extend ISC analysis to capture nonlinear and higher-order interactions in continuous fMRI responses. Specifically, given synchronized fMRI response frames in two brain regions $X$ and $Z$ across $K$ subjects $X_i, Z_i, i = 1, \ldots, K$ as random variables. We model the conditional mutual information $I(X_i; Z_j | i \neq j)$ as the MI form of pair-wise ISC analysis. By definition, $I(X_i; Z_j \mid i \neq j)$ first computes MI between activations $X_i$ and $Z_j$ from subjects $i$ and $j$ respectively, and then average across pairs of subjects $i \neq j$. It can be lower bounded using Eq. 7 by learning a $T_\theta(x, z)$ shared across all subject pairs.

**Dataset.** We collected an fMRI story comprehension dataset with 40 participants listening to four spoken stories and used it to study MI-based and correlation-based ISC on this dataset. The stories are renditions of "Pie Man"' (Pieman) and "Running from the Bronx" (Bronx) by Jim O'Grady [29, 30], "The Man Who Forgot Ray Bradbury" (Forgot) by Neil Gaiman [31], and "I Knew You Were Black' (Black)' by Carol Daniel [32]. Average story duration is 11 minutes. An

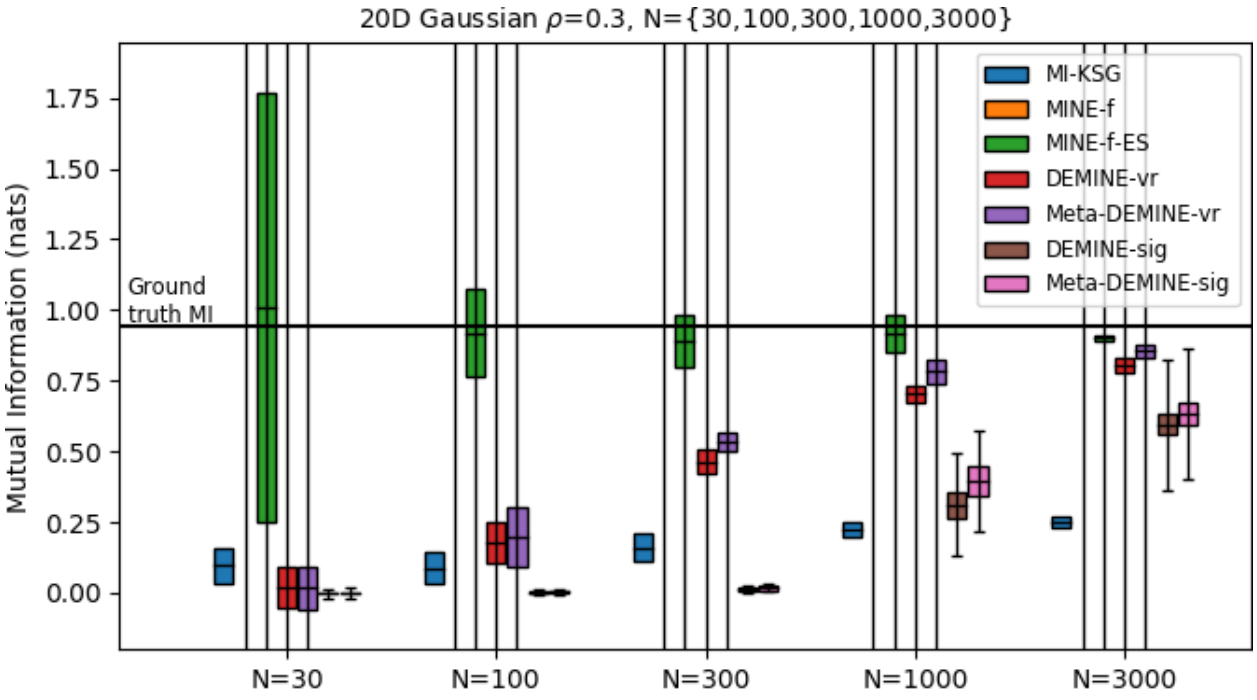



(a) 20D Gaussian dataset, $N = 300$ samples

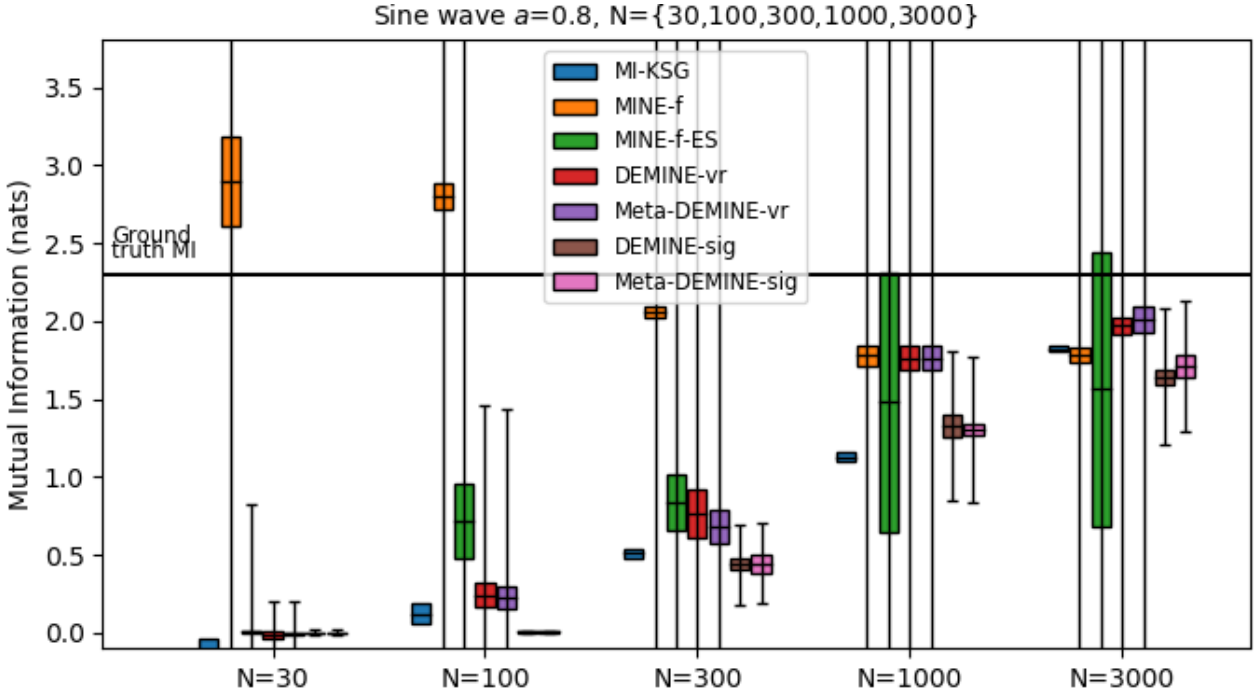



(b) 1D Gaussian dataset, $\rho = 0.8$

(Continuing on the next page)

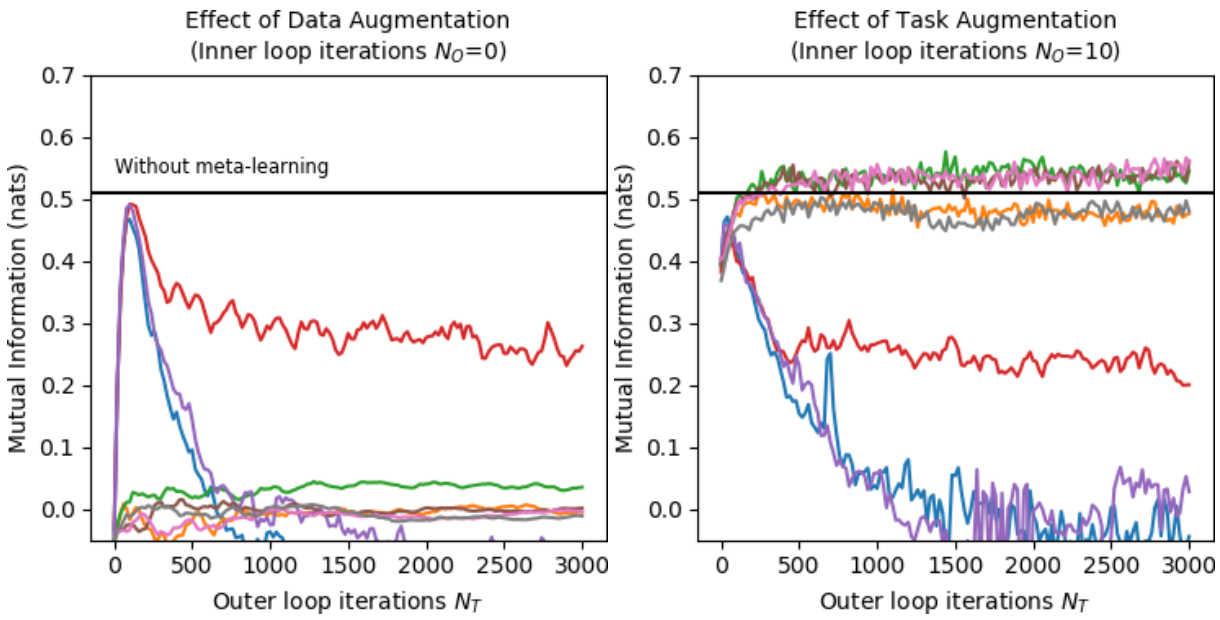



(c) 20D Gaussian dataset, $\rho = 0.3$



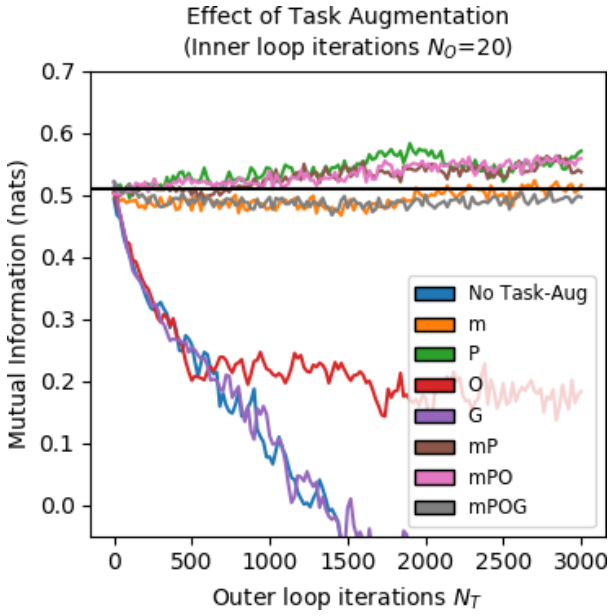

(d) Sine wave dataset, $a = 8\pi$

**Figure 3: Comparing MI Estimation performance of DEMINE and Meta-DEMINE with the KSG estimator [2] and MINE-f [1] on different datasets using varying number of samples. The bars show estimator mean and standard deviation averaged over 5 runs with different seeds. The errorbars show 95% confidence interval (not available for MI-KSG). The statistical significance focused variants DEMINE-sig and Meta-DEMINE-sig achieves the highest 95% confident MI estimation. Meta-DEMINE improves over DEMINE most of the time. Best viewed in color.**




(a) Meta-DEMINE-vr $N_O = 0$.

(b) Meta-DEMINE-vr $N_O = 10$.




(c) Meta-DEMINE-vr $N_O = 20$.

**Figure 4: To study the effect of task augmentation and number of adaptation steps, we run Meta-DEMINE-vr with different task augmentation modes and vary number of adaptation iterations $N_O \in \{0, 10, 20\}$ on Gaussian 20D, $\rho = 0.3$ dataset. Combinations of permutation and mirroring operations are effective in reducing overfitting and improving performance. Best viewed in color.**

Table 1: Number of HCP-MMP1 regions with significant pairwise correlation (r) and MI (DEMINE, Meta-DEMINE) during listening.

| Number of common regions with statistical significance | Pearson's r | DEMINE | Meta-DEMINE |
|---|---|---|---|
| Pearson's r | 37 | 24 | 23 |
| DEMINE | 24 | 28 | 26 |
| Meta-DEMINE | 23 | 26 | 29 |

fMRI frame with full brain coverage is captured at time of repetition (TR) 1 TR =1.5 seconds with 2.5mm isotropic spatial resolution. See Appendix A for additional data collection details and dataset statistics.We restricted our analysis to subsets of voxels defined using independent data from previous studies: functionally-defined masks of high ISC voxels (ISC; 3,800 voxels) and dorsal Default-Mode Network voxels (dDMN; 3,940 voxels) from [33], an anatomically-defined Gray Matter (GM) mask, as well as 180 Human Connectome Project Multi-Modal cortical Parcellation (HCP-MMP1) multimodal cortex parcels from [34]. All masks were defined in Montreal Neurological Institute (MNI) space.

**Implementation.** We compare MI-based ISC using DEMINE and Meta-DEMINE with correlation-based ISC using Pearson's correlation. DEMINE and Meta-DEMINE setup follows Section §4.1. The fMRI data were partitioned by subject into a train set of 20 subjects and a validation set of 20 different subjects. Residual 1D Convolutional Neural Networks (CNNs) are used instead of MLPs as the encoder for modeling temporal dependency. Network architecture is illustrated in Figure 5. For Pearson's correlation, high-dimensional signals are reshaped to 1D for correlation-based ISC analysis.

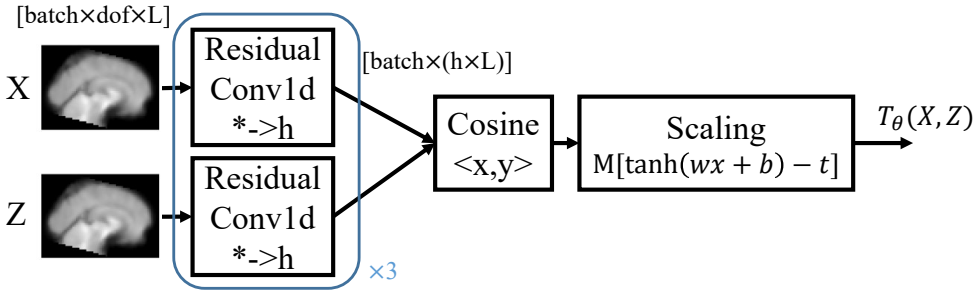



Figure 5: Network architecture for the fMRI experiments.

**Quantitative Results.** We first study that for the fine grained HCM-MMP1 brain regions, which of them have $p < 0.05$ statistically significant activities by MI and Pearson's correlation. Table 1 shows the result. Overall, more regions have statistically significant correlation than dependency. This is expected because correlation requires less data to detect. But Meta-DEMINE is able to find 6 brain regions that potentially have statistically significant dependency but lacks significant correlation. This shows that MI analysis can be used to complement correlation-based ISC analysis.

Table 2: Segment classification accuracy for NeuralMI versus Pearson's r in 1-vs-rest (1vR). All the results are averaging over other subjects. Abbreviations: P: Pieman; F: Forgot; Br: Bronx; Bk: Black, MI: Mutual Information.

| Classification Accuracy (%) | ISC Mask | | | | | dDMN Mask | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P | F | Br | Bk | MI | P | F | Br | Bk | MI |
| Chance | 3.7 | 1.8 | 2.6 | 1.9 | N/A | 3.7 | 1.8 | 2.6 | 1.9 | N/A |
| Pearson's r 1vR | 35.0 | 20.4 | 25.8 | 31.5 | N/A | 14.8 | 6.4 | **11.8** | 9.9 | N/A |
| DEMINE 1vR | 42.8 | 28.0 | 32.8 | 35.9 | 0.637 | **16.5** | **7.9** | 11.6 | **12.0** | **0.035** |
| Meta-DEMINE 1vR | **47.2** | **32.5** | **39.9** | **41.0** | **0.752** | 13.7 | **7.9** | 8.2 | 8.9 | 0.031 |

| Classification Accuracy (%) | GM Mask | | | | |
|---|---|---|---|---|---|
| | P | F | Br | Bk | MI |
| Chance | 3.7 | 1.8 | 2.6 | 1.9 | N/A |
| Pearson's r 1vR | 31.3 | 16.8 | 26.6 | 24.2 | N/A |
| DEMINE 1vR | **49.8** | **29.6** | **33.9** | **38.1** | **0.610** |
| Meta-DEMINE 1vR | - | - | - | - | - |

By considering temporal ISC over time, fMRI signals can be modeled with improved accuracy. In Table 2 we apply DEMINE-vr and Meta-DEMINE-vr with $L = 10$TRs (15s) sliding windows as random variables to study amount of information that can be extracted from ISC, dDMN and GM masks. We use between-subject time-segment classification (BSC) for evaluation [35, 36]. Each fMRI scan is divided into $K$ non-overlapping $L = 10$TRs time segments. The BSC task is one versus rest retrieval: retrieve the corresponding time segment $z$ of an individual given a group of time segments $x$ excluding that individual, measured by top-1 accuracy. For retrieval score, $T_\theta(X, Z)$ is used for DEMINE and Meta-DEMINE and $\rho(X, Z)$ is used for Pearson's correlation as a simple baseline. For DEMINE we report results on ISC and dDMN masks. For Meta-DEMINE we report results on ISC and dDMN masks as GM does not fit into the Graphics Processing Unit (GPU) memory that runs the algorithm due to the high memory consumption.

Using Convolutional Neural Networks (CNNs) as the encoders, DEMINE and Meta-DEMINE model the signal better and achieve higher accuracy. Also. Meta-DEMINE is able to extract 0.75 nats of MI from the ISC mask over 10TRs or 15s. We expect this to be improved by more samples and high frequency fMRI scans.

**Qualitative Results.** Figure 6 (top) visualizes voxels that are important to $T_\theta(x,\, z)$ of the DEMINE model using the magnitude variance of $T_\theta(x,\, z)$ with respect to voxels in $x$ and $z$ for the ISC, dDMN and GM masks. If a voxel has high gradient magnitude variance, it means that the voxel has high importance because its activation changes will induce large variations in $T_\theta(x,\, z)$. Results show that voxels from the auditory regions are most functionally important for perceiving the story stimulus.

Figure 6 (bottom) plots the $T_\theta(x,\, z)$ and inter-subject Pearson correlations over time for Pieman using the ISC mask and a sliding window size $L = 10$, using the one vs rest scores averaged over all subjects. DEMINE yields more distinctive peaks.

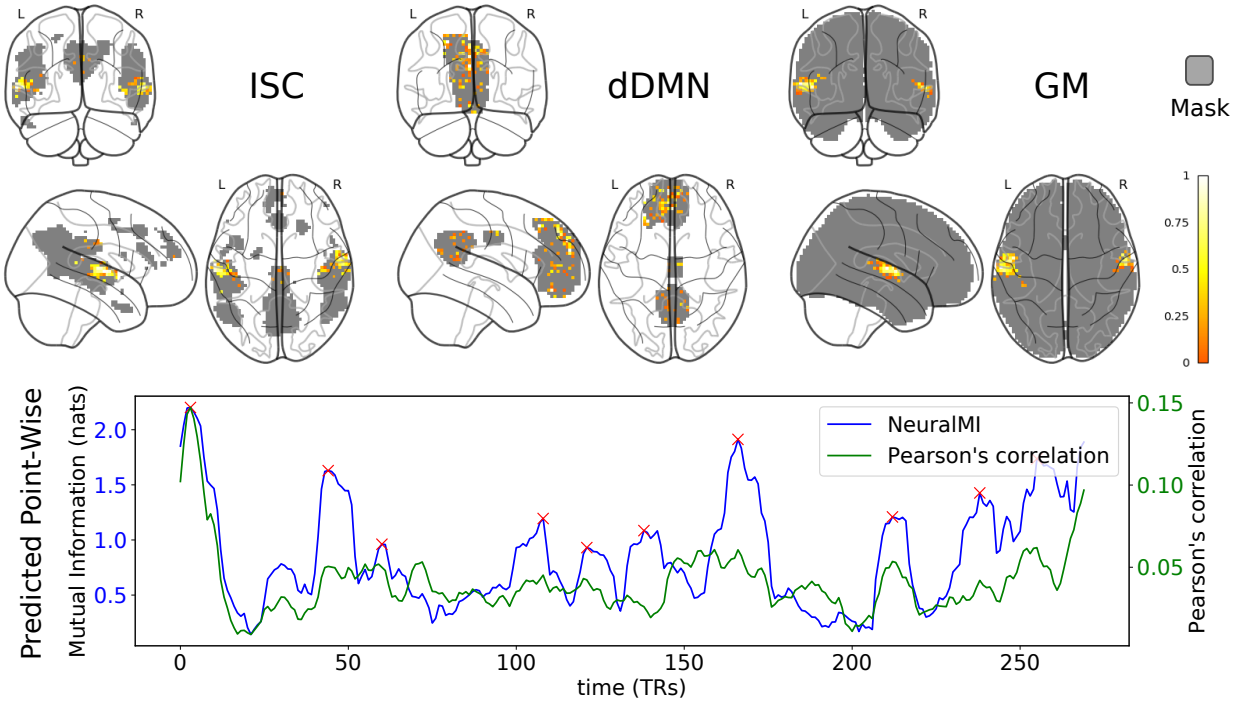



**Figure 6: Top: Top contributing voxels in the learned $T_\theta(X, Z)$ by gradient magnitude $\mathbb{E}_X(\frac{\partial T}{\partial X_i})^2$.** Auditory region is highlighted for ISC and GM masks (best in color). Bottom: Evaluation on the Pieman dataset using the ISC mask showing our approach $T_\theta(X, Z)$ versus Pearson correlation over time in the one versus rest case averaged over 20 test subjects.

We identify the peaks in DEMINE for Pieman (with Pearson correlations) over time, then locate the story transcriptions in the $L = 10$TRs (15 seconds) window corresponding to the peak:

- 4: "...toiled for The Ram, uh, Fordham University's student newspaper. And one day, I'm walking toward the campus center and out comes the elusive Dean McGowen, architect of a policy to replace traditionally ..."
- 45: "The Dean is covered with cream. So I give him a moment, then I say, 'Dean McGowen, would you care to comment on this latest attack?' And he says, 'Yes, I would care to comment. ..."
- 109: "...which makes no sense. Fordham was a Catholic school and we all thought Latin was classy so, that's what I used. And when I finished my story, I, I raced back to Dwyer and I showed it to him and he read it and he said ..."
- 122: "Few days later, I get a letter. I opened it up and it says, "Dear Jim, good story. Nice details. If you want to see me again in action, be on the steps of Duane Library ..."
- 139: "...out comes student body president, Sheila Biel. And now, Sheila Biel was different from the rest of us flannel-shirt wearing, part-time-job working, Fordham students. Sheila was..."

- 167: ”Pie Man emerged from behind a late night library drop box, made his delivery, and fled away, crying, “Ego sum non una bestia.” And that’s what I reported in my story. . . ”
- 213: ”. . . that there was a question about whether she even knew if I existed. So I saw her there and made a mental note to do nothing about it, and then I went to the bar and ordered a drink, and I felt a, a tap on my shoulder. I turned around, and it was her.. . . ”
- 239: ”And wasn’t I really Pie Man? Hadn’t I brought him into existence? Didn’t she only know about him because of me? But actually . . . ”
- 256: ”I said, “Yes, Angela, I am Pie Man.’ And she looked at me and she said, ‘Oh, good. I was hoping you’d say that . . . ”

We hypothesize that the scripts associated with the peaks may capture points when listeners pay more attention, resulting in the Signal-to-Noise Ratio (SNR) of fMRI scans being enhanced.

# 5.0 CONCLUSIONS

We illustrated that a predictive view of the MI lower bounds coupled with meta-learning results in data-efficient variational MI estimators, DEMINE and Meta-DEMINE, that are capable of performing statistical test of dependency. We benchmarked their effectiveness using synthetic datasets, and show that statistically significant dependency can be found using as low as 300 samples, and task augmentation reduces overfitting and improves generalization in meta-learning.

We successfully applied DEMINE dependency test to real world, data scarce, fMRI datasets. Our results suggest a greater avenue of using neural networks and meta-learning to improve MI analysis and applying neural network-based information theory tools to enhance the analysis of information processing in the brain.

Model-agnostic, high-confidence, MI lower bound estimation approaches – including *MINE*, DEMINE and Meta-DEMINE – are limited to estimating small MI lower bounds up to $O(\log N)$ as pointed out in [6], where $N$ is the number of samples. In real fMRI datasets, however, strong dependency is rare and existing MI estimation tools are limited more by their ability to accurately characterize the dependency. Nevertheless, when quantitatively measuring strong dependency, alternatives to MI – cross-entropy [6] or model-based quantities such as correlation or Canonical Correlation Analysis (CCA) – may be measured with high confidence.

# 6.0 REFERENCES


[1] Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Hjelm, D., and Courville, A. "Mutual information neural estimation". In *International Conference on Machine Learning*, pp. 530–539, 2018.

[2] Kraskov, A., Stogbauer, H., and Grassberger, P. "Estimating mutual information". *Physical review E*, 2004.

[3] Gao, W., Kannan, S., Oh, S., and Viswanath, P. "Estimating mutual information for discrete-continuous mixtures". In *Advances in Neural Information Processing Systems*, pp. 5986–5997, 2017.

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

[5] Ahmad, I. and Lin, P.-E. "A nonparametric estimation of the entropy for absolutely continuous distributions (corresp.)". *IEEE Transactions on Information Theory* ***22*****(3)**, 372–375, 1976.

[6] McAllester, D. and Statos, K. "Formal limitations on the measurement of mutual information". *arXiv preprint arXiv:1811.04251*, 2018.

[7] Agakov, D. B. F. "The IM algorithm: a variational approach to information maximization". *Advances in Neural Information Processing Systems* **16**, 201, 2004.

[8] Poole, B., Ozair, S., van den Oord, A., Alemi, A. A., and Tucker, G. "On variational lower bounds of mutual information". In *Bayesian Deep Learning Workshop, NeurIPSW*, 2018.

[9] Maclaurin, D., Duvenaud, D., and Adams, R. "Gradient-based hyperparameter optimization through reversible learning". In *International Conference on Machine Learning*, pp. 2113–2122, 2015.

[10] Pham, H., Guan, M., Zoph, B., Le, Q., and Dean, J. "Efficient neural architecture search via parameter sharing". In *International Conference on Machine Learning*, pp. 4092–4101, 2018.

[11] Finn, C., Abbeel, P., and Levine, S. "Model-agnostic meta-learning for fast adaptation of deep networks". In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 1126–1135, 2017.

[12] Finn, C., Xu, K., and Levine, S. "Probabilistic model-agnostic meta-learning". In *Advances in Neural Information Processing Systems*, pp. 9537–9548, 2018.

[13] Kim, T., Yoon, J., Dia, O., Kim, S., Bengio, Y., and Ahn, S. "Bayesian model-agnostic meta-learning". *arXiv preprint arXiv:1806.03836*, 2018.

[14] Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. "Matching networks for one shot learning". In *Advances in neural information processing systems*, pp. 3630–3638, 2016.

[15] Snell, J., Swersky, K., and Zemel, R. "Prototypical networks for few-shot learning". In *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017.

[16] Finn, C., Yu, T., Zhang, T., Abbeel, P., and Levine, S. "One-shot visual imitation learning via meta-learning". In *Conference on Robot Learning*, pp. 357–368, 2017.

[17] Glorot, X. and Bengio, Y. "Understanding the difficulty of training deep feedforward neural networks". In *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10). Society for Artificial Intelligence and Statistics*, 2010.

[18] Hoeffding, W. "Probability inequalities for sums of bounded random variables". In *The Collected Works of Wassily Hoeffding*, pp. 409–426. Springer, 1994.

[19] Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. "Understanding deep learning requires rethinking generalization". *arXiv preprint arXiv:1611.03530*, 2016.

[20] Kingma, D. P. and Ba, J. "Adam: A method for stochastic optimization". *arXiv preprint arXiv:1412.6980*, 2014.

[21] Salimans, T., Ho, J., Chen, X., Sidor, S., and Sutskever, I. "Evolution strategies as a scalable alternative to reinforcement learning". *arXiv preprint arXiv:1703.03864*, 2017.

[22] Sehnke, F., Osendorfer, C., Rückstieß, T., Graves, A., Peters, J., and Schmidhuber, J. "Parameter-exploring policy gradients". *Neural Networks* ***23*****(4)**, 551–559, 2010.

[23] Bergstra, J., Yamins, D., and Cox, D. D. "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures", 2013.

[24] Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., and Keysers, C. "Brain-to-brain coupling: a mechanism for creating and sharing a social world". *Trends in cognitive sciences* ***16*****(2)**, 114–121, 2012.

[25] Stephens, G. J., Silbert, L. J., and Hasson, U. "Speaker–listener neural coupling underlies successful communication". *Proceedings of the National Academy of Sciences* ***107*****(32)**, 14425–14430, 2010.

[26] Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., and Malach, R. "Intersubject synchronization of cortical activity during natural vision". *Science* ***303*****(5664)**, 1634–1640, 2004.

[27] Schippers, M. B., Roebroeck, A., Renken, R., Nanetti, L., and Keysers, C. "Mapping the information flow from one brain to another during gestural communication". *Proceedings of the National Academy of Sciences*, 201001791, 2010.

[28] Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., and Hasson, U. "Coupled neural systems underlie the production and comprehension of naturalistic narrative speech". *Proceedings of the National Academy of Sciences* ***111*****(43)**, E4687–E4696, 2014.

[29] O'Grady, J. "Pie Man". `https://themoth.org/stories/pie-man`, 2018a. Accessed: 2018-10-12.

[30] O'Grady, J. "Running from the Bronx". https://soundcloud.com/the-story-collider/jim-ogrady-running-from-the, 2018b. Accessed: 2018-10-12.

[31] Gaiman, N. "The man who forgot ray bradbury". https://soundcloud.com/neilgaiman/the-man-who-forgot-ray-bradbury, 2018. Accessed: 2018-10-12.

[32] Daniel, C. "I knew you were black". https://themoth.org/stories/i-knew-you-were-black, 2018. Accessed: 2018-10-12.

[33] Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., and Hasson, U. "Dynamic reconfiguration of the default mode network during narrative comprehension". *Nature Communications* **7**, 12141, 2016.

[34] Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., et al. "A multi-modal parcellation of human cerebral cortex". *Nature* ***536*****(7615)**, 171, 2016.

[35] Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Hanke, M., and Ramadge, P. J. "A common, high-dimensional model of the representational space in human ventral temporal cortex". *Neuron* ***72*****(2)**, 404–416, 2011.

[36] Guntupalli, J. S., Hanke, M., Halchenko, Y. O., Connolly, A. C., Ramadge, P. J., and Haxby, J. V. "A model of representational spaces in human cortex". *Cerebral Cortex* ***26*****(6)**, 2919–2934, 2016.

[37] Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., et al. "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments". *Scientific Data* **3**, 160044, 2016.

[38] Esteban, O., Markiewicz, C., Blair, R. W., Moodie, C., Isik, A. I., Erramuzpe Aliaga, A., Kent, J., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S., Wright, J., Durnez, J., Poldrack, R., and Gorgolewski, K. J. "FMRIPrep: a robust preprocessing pipeline for functional MRI". *bioRxiv*, 2018.

[39] Cox, R. W. "AFNI: software for analysis and visualization of functional magnetic resonance neuroimages". *Computers and Biomedical research* ***29*****(3)**, 162–173, 1996.

[40] Gorgolewski, K., Burns, C., Madison, C., Clark, D., Halchenko, Y., Waskom, M., and Ghosh, S. "Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python". *Frontiers in Neuroinformatics* **5**, 13, 2011.

[41] Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. "N4itk: improved n3 bias correction". *IEEE Transactions on Medical Imaging* ***29*****(6)**, 1310–1320, June 2010.

[42] Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain". *Medical Image Analysis* ***12*****(1)**, 26–41, 2008.

[43] Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C., and Collins, D. "Unbiased nonlinear average age-appropriate brain templates from birth to adulthood". *NeuroImage* **(47)**, S102, 2009.

[44] Zhang, Y., Brady, M., and Smith, S. "Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm". *IEEE Transactions on Medical Imaging* ***20*****(1)**, 45–57, 2001.

[45] Jenkinson, M., Bannister, P., Brady, M., and Smith, S. "Improved optimization for the robust and accurate linear registration and motion correction of brain images". *NeuroImage* ***17*****(2)**, 825–841, 2002.

[46] Wang, S., Peterson, D. J., Gatenby, J. C., Li, W., Grabowski, T. J., and Madhyastha, T. M. "Evaluation of field map and nonlinear registration methods for correction of susceptibility artifacts in diffusion mri". *Frontiers in Neuroinformatics* **11**, 17, 2017.

[47] Treiber, J. M., White, N. S., Steed, T. C., Bartsch, H., Holland, D., Farid, N., McDonald, C. R., Carter, B. S., Dale, A. M., and Chen, C. C. "Characterization and correction of geometric distortions in 814 diffusion weighted images". *PLOS ONE* ***11*****(3)**, e0152472, 2016.

[48] Greve, D. N. and Fischl, B. "Accurate and robust brain image alignment using boundary-based registration". *NeuroImage* ***48*****(1)**, 63–72, 2009.

[49] Behzadi, Y., Restom, K., Liau, J., and Liu, T. T. "A component based noise correction method (CompCor) for BOLD and perfusion based fMRI". *NeuroImage* ***37*****(1)**, 90–101, 2007.

[50] Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. "Methods to detect, characterize, and remove motion artifact in resting state fMRI". *NeuroImage* **84**, 320–341, 2014.

# APPENDIX A. Additional Details of the fMRI Dataset

The dataset we collected contains 40 participants (mean age = 23.3 years, standard deviation = 8.9, range: 18–53; 27 female) recruited to listen to four spoken stories[5, 6]. The stories were renditions of "Pie Man" and "Running from the Bronx" by Jim O'Grady [29, 30], "The Man Who Forgot Ray Bradbury" by Neil Gaiman [31], and "I Knew You Were Black" by Carol Daniel [32]; story durations were 7, 9, 14, and 13 minutes, respectively. After scanning, participants completed a questionnaire comprising 25–30 questions per story intended to measure narrative comprehension. The questionnaires included multiple choice, True/False, and fill-in-the-blank questions, as well as four additional subjective ratings per story. Functional and structural images were acquired using a 3T Siemens Prisma with a 64-channel head coil. Briefly, functional images were acquired in an interleaved fashion using gradient-echo echo-planar imaging with a multiband acceleration factor of 3 (TR/TE = 1500/31 ms where TE stands for "echo time", resolution = 2.5 mm isotropic voxels, full brain coverage).

All fMRI data were formatted according to the Brain Imaging Data Structure (BIDS) standard [37] and preprocessed using the fMRIPrep library [38]. Functional data were corrected for slice timing, head motion, and susceptibility distortion, and normalized to MNI space using nonlinear registration. Nuisance variables comprising head motion parameters, framewise displacement, linear and quadratic trends, sine/cosine bases for high-pass filtering (0.007 Hz), and six principal component time series from cerebrospinal fluid (CSF) and white matter (WM) were regressed out of the signal using the Analysis of Functional NeuroImages (AFNI) software suite [39].

The fMRI data comprise $\mathcal{X} \in \mathbb{R}^{V_i \times T}$ for each subject, where $V_i$ represents the flattened and masked voxel space and $T$represents the number of samples (in TRs) during auditory stimulus presentation.

**Additional Details on Dataset Collection** Functional and structural images were acquired using a 3T Siemens Magnetom Prisma with a 64-channel head coil. Functional, blood-oxygenation-level-dependent (BOLD) images were acquired in an interleaved fashion using gradient-echo echo-planar imaging with pre-scan normalization, fat suppression, a multiband acceleration factor of 3, and no in-plane acceleration: TR/TE = 1500/31 ms, flip angle = 67°, bandwidth = 2480 hz per pixel, resolution = 2.5 $mm^3$ isotropic voxels, matrix size = 96 x 96, Fied of view (FoV) = 240 x 240 mm, 48 axial slices with roughly full brain coverage and no gap, anterior–posterior phase encoding. At the beginning of each scanning session, a T1-weighted structural scan (where T1 stands for "longitudinal relaxation time"), was acquired using a high-resolution single-shot Magnetization-Prepared 180 degrees radio-frequency pulses and RApid Gradient-Echo (MPRAGE) sequence with an in-plane acceleration factor of 2 using GeneRalized Autocalibrating Partial Parallel Acquisition (GRAPPA): TR/TE/TI = 2530/3.3/1100 ms where TI stands for inversion time, flip angle = 7°, resolution = 1.0 x 1.0 x 1.0 mm voxels, matrix size = 256 x 256, FoV = 256 x 256 x 176 mm, 176 sagittal slices, ascending acquisition, anterior–posterior phase encoding, no fat suppression, 5 min 53 s total acquisition time. At the end of each scanning session a T2-weighted (where T2 stands for

[5]Two of the stories were told by a professional storyteller undergoing an fMRI scan; however, fMRI data for the speaker were not analyzed for the present work due to the head motion induced by speech production.

[6]The study was conducted in compliance with the Institutional Review Board of the University

"transverse relaxation time") structural scan was acquired using the same acquisition parameters and geometry as the T1-weighted structural image: TR/TE = 3200/428 ms, 4 minutes 40 seconds total acquisition time. A field map was acquired at the beginning of each scanning session, but was not used in subsequent analyses.

**Additional Details on Dataset Preprocessing** Preprocessing was performed using the fMRIPrep library[7] [38], a Nipype library[8] [40] based tool. T1-weighted images were corrected for intensity non-uniformity using the N4 bias field correction algorithm [41] and skull-stripped using Advanced Normalization Tools (ANTs) [42]. Nonlinear spatial normalization to the International Consortium for Brain Mapping (ICBM) 152 Nonlinear Asymmetrical template version 2009c [43] was performed using ANTs. Brain tissue segmentation cerebrospinal fluid, white matter, and gray matter was was performed using FSL library's[9] FAST tool [44]. Functional images were slice timing corrected using AFNI software's 3dTshift [39] and corrected for head motion using FSL library's MCFLIRT tool [45]. "Fieldmap-less" distortion correction was performed by co-registering each subject's functional image to that subject's intensity-inverted T1-weighted image [46] constrained with an average field map template [47]. This was followed by co-registration to the corresponding T1-weighted image using FreeSurfer software's[10] boundary-based registration [48] with 9 degrees of freedom. Motion correcting transformations, field distortion correcting warp, BOLD-to-T1 transformation and T1-to-template (MNI) warp were concatenated and applied in a single step with Lanczos interpolation using ANTs. Physiological noise regressors were extracted applying "a Component Based Noise Correction Method" aCompCor [49]. Six principal component time series were calculated within the intersection of the subcortical mask and the union of CSF and WM masks calculated in T1w (T1 for white matter) space, after their projection to the native space of each functional run. Framewise displacement [50] was calculated for each functional run. Functional images were downsampled to 3 mm resolution. Nuisance variables comprising six head motion parameters (and their derivatives), framewise displacement, linear and quadratic trends, sine/cosine bases for high-pass filtering (0.007 Hz cutoff), and six principal component time series from an anatomically-defined mask of cerebrospinal fluid and white matter were regressed out of the signal using AFNI's 3dTproject [39]. Functional response time series were z-scored for each voxel.

[7] https://github.com/poldracklab/fmriprep

[8] https://github.com/nipy/nipype

[9] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSL

[10] https://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferWiki

# LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

| ACRONYM | DESCRIPTION |
|---|---|
| 1D | 1-dimensional |
| 20D | 20-dimensional |
| aCompCor | A component based noise correction method |
| AFNI | Analysis of Functional NeuroImages |
| AFRL | Air Force Research Laboratory |
| ANTs | Advanced Normalization Tools |
| BIDS | Brain Imaging Data Structure |
| Black | "I Knew You Were Black", story by Carol Daniel |
| BOLD | Blood-oxygenation-level-dependent |
| BPTT | Backpropagation through time |
| Bronx | "Running from the Bronx", story by Jim O'Grady |
| BSC | Between-subject time-segment classification |
| CCA | Canonical Correlation Analysis |
| CNNs | Convolutional Neural Networks |
| CSF | Cerebrospinal fluid |
| DARPA | Defense Advanced Research Projects Agency |
| dDMN | Dorsal Default-Mode Network |
| DEMINE | Data-Efficient MINE Estimator |
| DEMINE-sig | DEMINE for statistical significance |
| DEMINE-vr | DEMINE with variance reduction |
| DoD | Department of Defense |
| ES | Evolution Strategies |
| fMRI | functional Magnetic Resonance Imaging |
| Forgot | "The Man Who Forgot Ray Bradbury", story by Neil Gaiman |
| FoV | Field of view |
| GANs | Generative Adversarial Networks |
| GM | Gray matter |
| GPU | Graphics Processing Unit |
| GRAPPA | GeneRalized Autocalibrating Partial Parallel Acquisition |
| HCP-MMP1 | Human Connectome Project Multi-Modal cortical Parcellation, version 1 |
| *i.i.d.* | Independent and identically distributed |
| ICBM | International Consortium for Brain Mapping |

| ACRONYM | DESCRIPTION |
|---|---|
| IM | Information Maximization |
| ISC | Inter-Subject Correlation |
| k-NN | k-Nearest Neighbors |
| KL | Kullback-Leibler |
| KSG | Kraskov-Stogbauer-Grassberger |
| MAML | Model-Agnostic Meta-Learning |
| Meta-DEMINE | Meta-learned Data-Efficient MINE Estimator |
| Meta-DEMINE-sig | Meta-DEMINE for statistical significance |
| Meta-DEMINE-vr | Meta-DEMINE with variance reduction |
| MI | Mutual Information |
| MI-KSG | The Kraskov-Stogbauer-Grassberger mutual information estimator |
| MINE | Mutual Information Neural Estimation/Estimator |
| MINE-f | MINE with f-divergence |
| MINE-f ES | MINE-f with early stopping |
| MLPs | Multi-layer perceptrons |
| MNI | Montreal Neurological Institute |
| MPRAGE | Magnetization-Prepared 180 degrees radio-frequency pulses and RApid Gradient-Echo |
| N4 | Improved N3 (nonparametric nonuniform intensity normalization) algorithm |
| PEPG | Parameter-Exploring Policy Gradient |
| Pieman | "Pie Man", story by Jim O'Grady |
| ReLU | Rectified linear units |
| ROIs | Regions of Interest |
| SGD | Stochastic Gradient Descent |
| SNR | Signal-to-Noise Ratio |
| T1 | The longitudinal relaxation time |
| T1w | T1 for white matter |
| T2 | The transverse relaxation time |
| TE | Echo time |
| The IM algorithm | The Information Maximization algorithm |
| The KSG estimator | The Kraskov-Stogbauer-Grassberger estimator |
| The TCPC estimator | The Contrastive Predictive Coding estimator with bias-variance tradeoff |
| TI | The inversion time |
| TR | Time of repetition |
| VAEs | Variational Autoencoders |
| WM | White matter |