A Bayesian Perspective On Training Data Attribution: Supplementary Material

Anonymous Author(s)
Affiliation
Address
email

A Model training details

We provide the source code anonymously for double-blind review at `https://anonymous.4open.science/r/bayesian-tda-14D6/` and in the attached zip file. We plan to make the repository public upon acceptance. All experiments were run on a single Nvidia 2080ti GPU.

A.1 Data sampling

We use subsampled versions of the openly available MNIST [1] and CIFAR10 [2] datasets. For this, we first add an index which we use for randomly sampling a fixed number of images from each class. Table 1 includes the dataset sizes of the different experiments.

A.2 Model training

The CNN has two convolutional layers followed by two fully connected linear layers, with GeLU activation. We use the Adam optimizer with a learning rate of 0.001 and a weight decay of 0.005. We use the cross-entropy loss and train the model for 15 epochs on MNIST3 and for 30 epochs on CIFAR10 with a batch size of 32. For training the ViT with LoRA, we use the `peft` [3] and HuggingFace `transformers` library [4]. We start from the pretrained model checkpoint of [5] and finetune the LoRA model with the same hyperparameters as the CNN.

An overview of the predictive performance measured in accuracy on the subsampled training and test sets is provided in Table 1.

Hint for reproducibility. In particular, we use `CrossEntropyLoss(reduction='none')` during model training and also update this in the ViT training script `modeling_vit.py`. This is important for the LOO experiments, where we exclude a sample z_j from contributing to the training by zeroing out the loss.

B Additional experimental results

We study the reliability of TDA estimates and values through a hypothesis test, where we report the p-values as an indication of the statistical significance of the TDA estimate. In this appendix, we provide a complete overview of the analyses of p-values and correlations between different TDA methods.

B.1 DE vs. DE+SWA

In our work, we sample trained models θ sampled from the posterior $p(\theta|\mathcal{D})$. Concretely, we train the model across 10 different random seeds and record the checkpoints after each of the last five epochs

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Table 1: Predictive performance at 95% CI across 10 runs (computed as 1.96*SE)

Experiment						
Model	Data	Randomness	$\lvert\mathcal{D}_{\text{train}}\rvert$	$\lvert\mathcal{D}_{\text{test}}\rvert$	$\text{Accuracy}_{\text{train}}$	$\text{Accuracy}_{\text{test}}$
CNN	MNIST3	SWA+DE-Init	30	900	0.987±0.010	0.953±0.003
CNN	MNIST3	SWA+DE-Init	60	900	0.985±0.007	0.970±0.004
CNN	MNIST3	SWA+DE-Init	150	900	0.998±0.004	0.970±0.003
CNN	CIFAR10	SWA+DE-Init	100	500	0.989±0.020	0.260±0.010
CNN	CIFAR10	SWA+DE-Init	200	500	1.000±0.000	0.328±0.007
CNN	CIFAR10	SWA+DE-Init	500	500	0.992±0.014	0.377±0.012
CNN	MNIST3	SWA+DE-Batch	150	900	0.993±0.002	0.975±0.002
CNN	CIFAR10	SWA+DE-Batch	500	500	0.991±0.010	0.364±0.003
ViT+LoRA	MNIST3	SWA+DE-Batch	150	900	0.945±0.008	0.935±0.005
ViT+LoRA	CIFAR10	SWA+DE-Batch	500	500	0.934±0.006	0.892±0.008

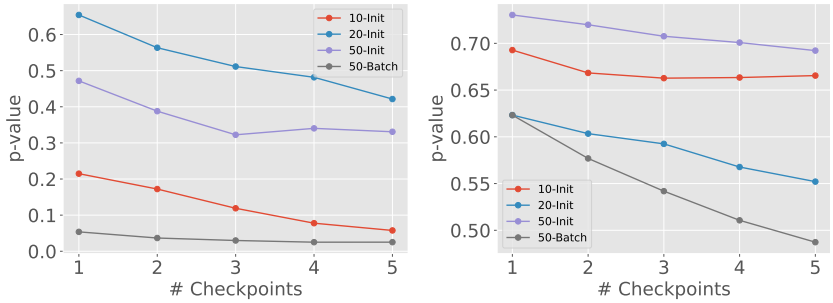

Figure 1: P-values of ground-truth TDA (LOO) with increasing number of SWA samples (i.e., number of model checkpoints used as samples of θ) for the CNN trained on MNIST3 and CIFAR10.

of training. Each of these models represents a sample. We test how the stability of the TDA values $\tau(z_j, z)$ (LOO) behaves when we ensemble different numbers of checkpoints by investigating the mean p-values across all train-test pairs (z_j, z). Figure 1 shows that higher numbers of samples θ increase stability, therefore we use all available samples in our subsequent analyses.

B.2 All results: Stability of TDA values and estimates

Table 2 presents the complete table of p-values of all tested TDA methods across all experiments of our work. Below, we display the histograms of p-values per experiment corresponding to each line in the table captioned with the experiment ID. The histograms show that low-noise train-test pairs (z_j, z) are present in all experiments involving the CNN model (i.e., experiments 1-8), where the number of pairs varies. Generally, we observe that there is no connection between the size of the dataset and the distribution of p-values. Furthermore, we notice that fixing the model initialisation (i.e., randomness induced by SWA+DE-Batch) increases the number of stable train-test pairs (cf. experiment 3 to 7, 4 to 8). However, in the case of the ViT experiments, stable train-test pairs are practically non-existent which shows that model complexity affects the stability of TDA.

B.3 All results: Correlation analysis

In the main body of this paper, we report the Pearson and Spearman correlation matrices for experiment 3 (CNN trained on MNIST3 with 50 samples per class and randomness induced by SWA+DE-Init). This section presents the complete overview of correlations between TDA methods across all experiments in Figures 3 - 12. We note that the observations and analyses in the main paper hold across experiments.

Table 2: Complete list of mean p-values of TDA values for all experiments.

Experiment									
ID	Model	Data	Randomness	$\lvert\mathcal{D}_{\text{train}}\rvert$	LOO	ATS	IF	GD	GC
1	CNN	MNIST3	SWA+DE-Init	30	0.058	0.088	0.111	0.110	0.020
2	CNN	MNIST3	SWA+DE-Init	60	0.421	0.148	0.251	0.253	0.002
3	CNN	MNIST3	SWA+DE-Init	150	0.331	0.254	0.352	0.363	0.003
4	CNN	CIFAR10	SWA+DE-Init	100	0.665	0.424	0.607	0.608	0.352
5	CNN	CIFAR10	SWA+DE-Init	200	0.552	0.397	0.450	0.452	0.399
6	CNN	CIFAR10	SWA+DE-Init	500	0.692	0.438	0.575	0.587	0.356
7	CNN	MNIST3	SWA+DE-Batch	150	0.025	0.039	0.000	0.000	0.000
8	CNN	CIFAR10	SWA+DE-Batch	500	0.623	0.374	0.535	0.534	0.314
9	ViT+LoRA	MNIST3	SWA+DE-Batch	150	0.786	0.573	0.369	0.365	0.093
10	ViT+LoRA	CIFAR10	SWA+DE-Batch	500	0.777	0.766	0.686	0.686	0.522

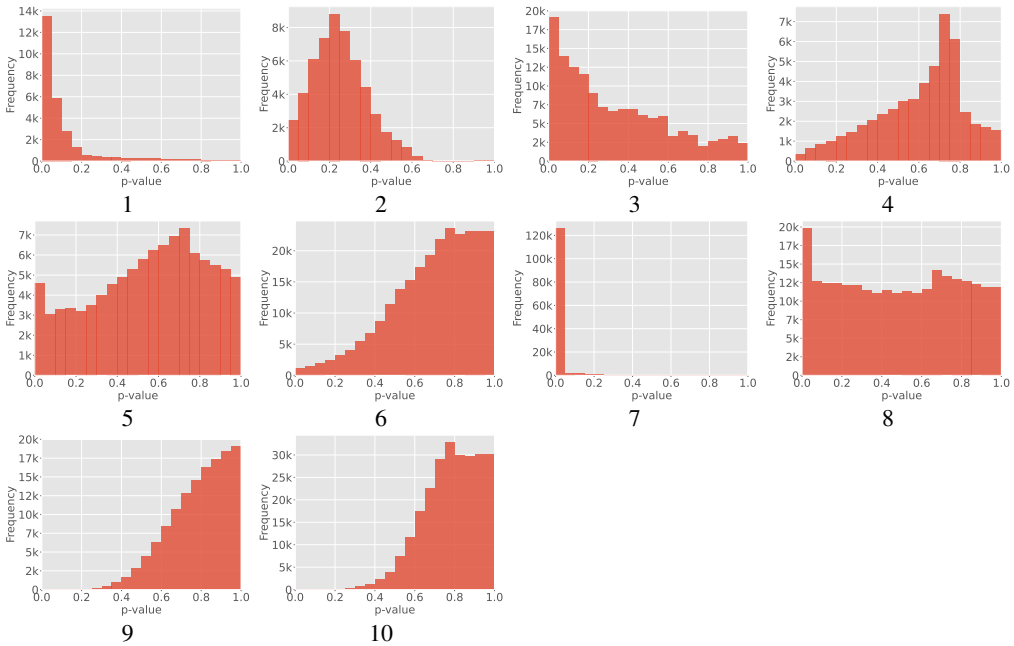

Figure 2: Distribution of p-values for ground-truth TDA (LOO) for all experiments (IDs corresponding to IDs in Table 2).

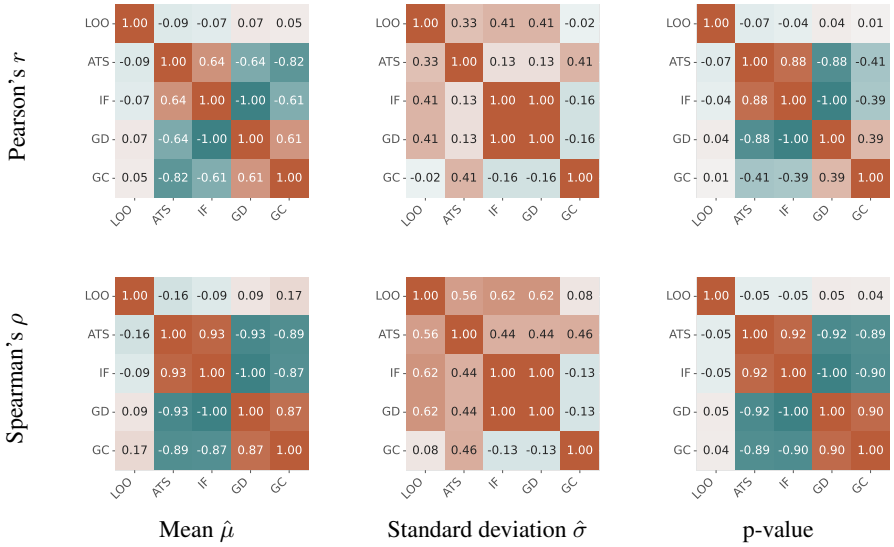

Figure 3: **Experiment 1 (cf. Table 2):** Pearson and Spearman correlation coefficients among ground-truth TDA and approximate TDA methods. We show correlations for TDA mean $\hat{\mu}$, TDA standard deviation $\hat{\sigma}$, and TDA p-values.

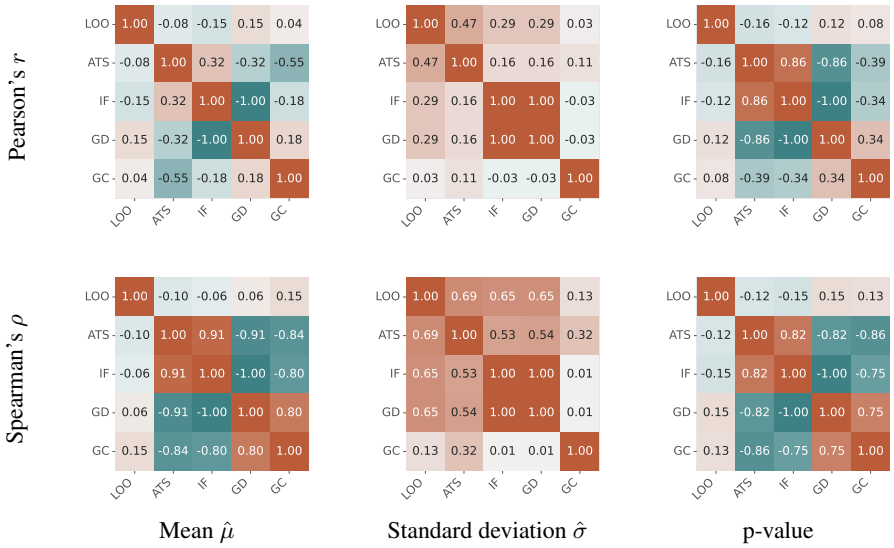

Figure 4: **Experiment 2 (cf. Table 2):** Pearson and Spearman correlation coefficients among ground-truth TDA and approximate TDA methods. We show correlations for TDA mean $\hat{\mu}$, TDA standard deviation $\hat{\sigma}$, and TDA p-values.

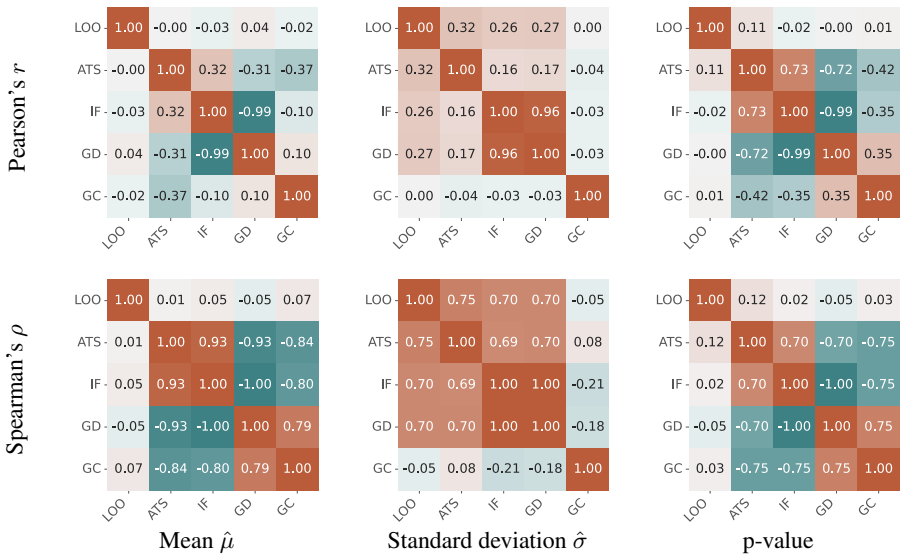

Figure 5: **Experiment 3 (cf. Table 2):** Pearson and Spearman correlation coefficients among ground-truth TDA and approximate TDA methods. We show correlations for TDA mean $\hat{\mu}$, TDA standard deviation $\hat{\sigma}$, and TDA p-values.

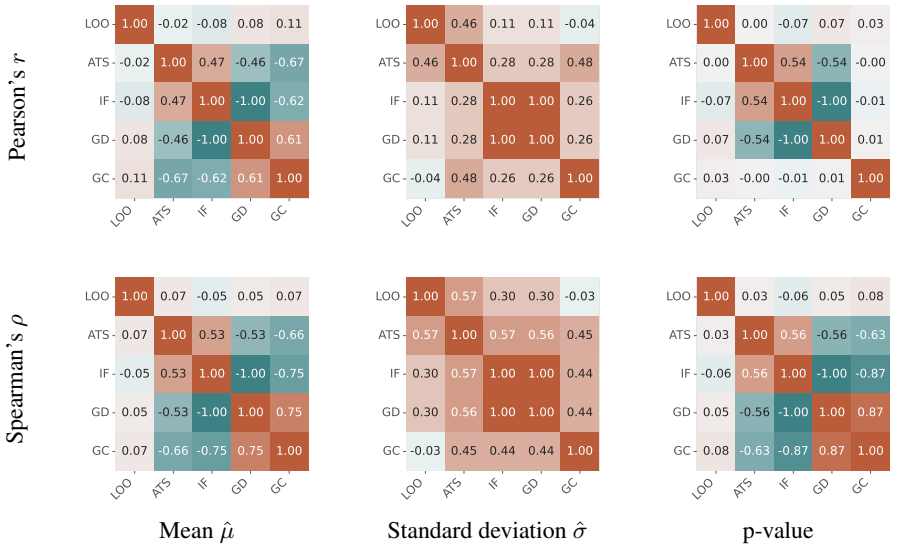

Figure 6: **Experiment 4 (cf. Table 2):** Pearson and Spearman correlation coefficients among ground-truth TDA and approximate TDA methods. We show correlations for TDA mean $\hat{\mu}$, TDA standard deviation $\hat{\sigma}$, and TDA p-values.

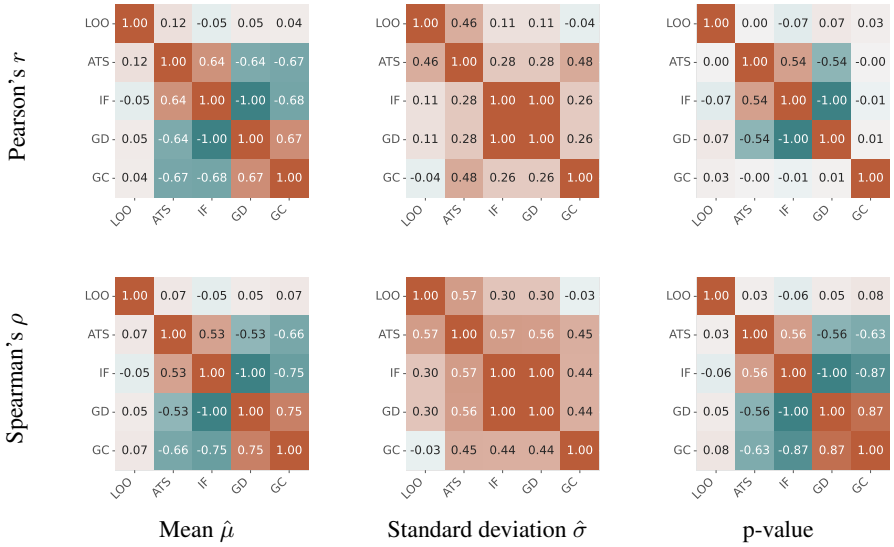

Figure 7: **Experiment 5 (cf. Table 2):** Pearson and Spearman correlation coefficients among ground-truth TDA and approximate TDA methods. We show correlations for TDA mean $\hat{\mu}$, TDA standard deviation $\hat{\sigma}$, and TDA p-values.

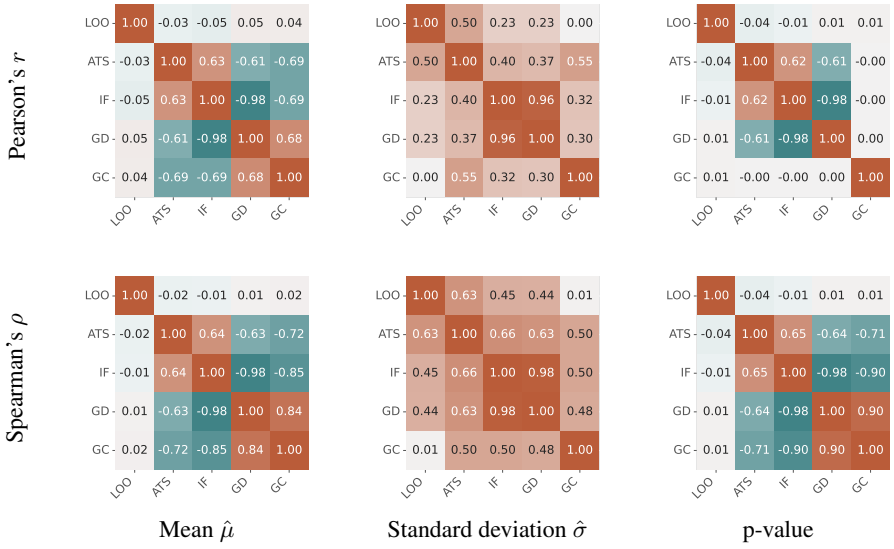

Figure 8: **Experiment 6 (cf. Table 2):** Pearson and Spearman correlation coefficients among ground-truth TDA and approximate TDA methods. We show correlations for TDA mean $\hat{\mu}$, TDA standard deviation $\hat{\sigma}$, and TDA p-values.

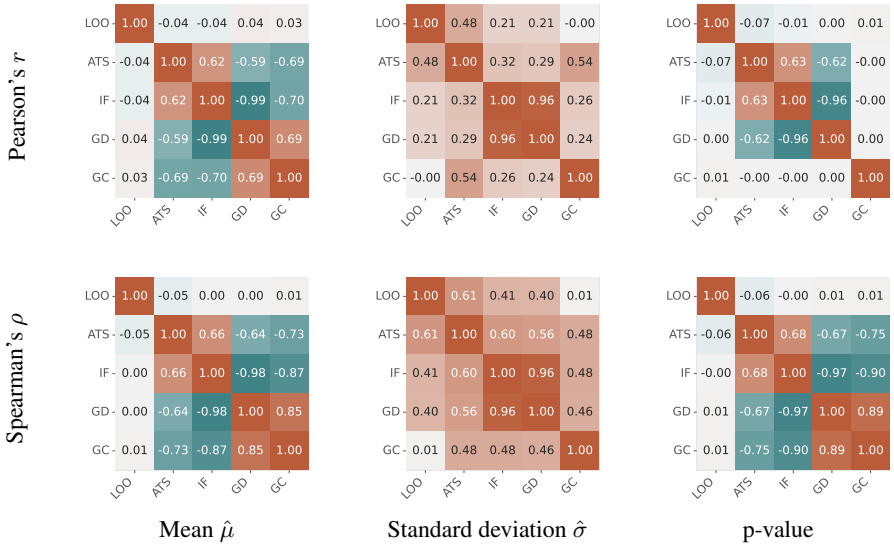

Figure 9: **Experiment 7 (cf. Table 2):** Pearson and Spearman correlation coefficients among ground-truth TDA and approximate TDA methods. We show correlations for TDA mean $\hat{\mu}$, TDA standard deviation $\hat{\sigma}$, and TDA p-values.

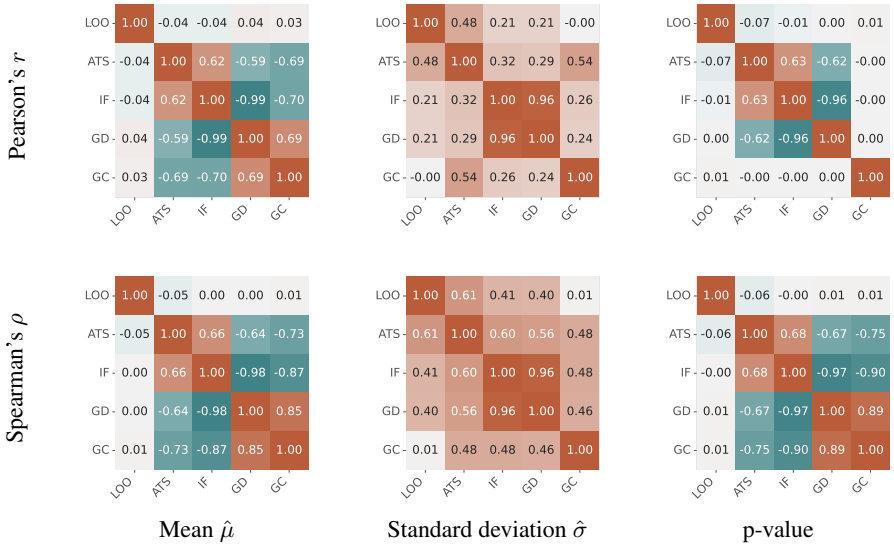

Figure 10: **Experiment 8 (cf. Table 2):** Pearson and Spearman correlation coefficients among ground-truth TDA and approximate TDA methods. We show correlations for TDA mean $\hat{\mu}$, TDA standard deviation $\hat{\sigma}$, and TDA p-values.

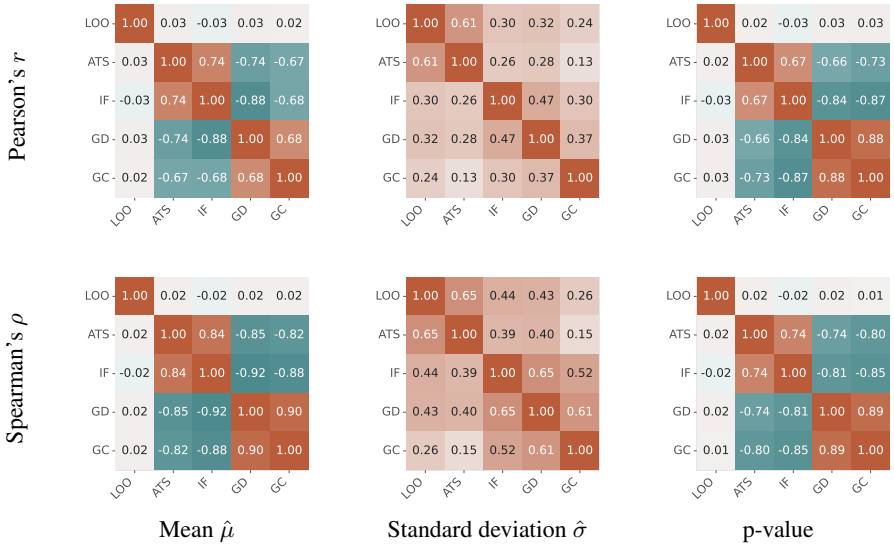

Figure 11: **Experiment 9 (cf. Table 2):** Pearson and Spearman correlation coefficients among ground-truth TDA and approximate TDA methods. We show correlations for TDA mean $\hat{\mu}$, TDA standard deviation $\hat{\sigma}$, and TDA p-values.

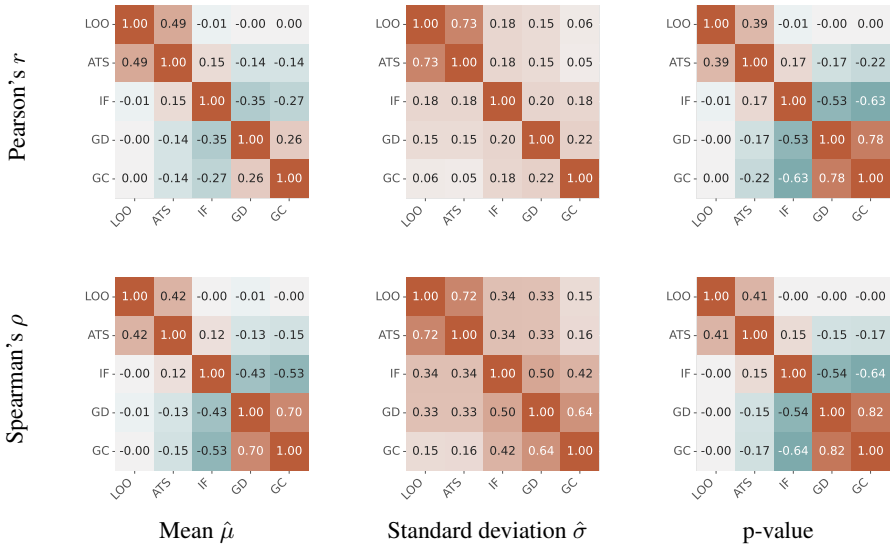

Figure 12: **Experiment 10 (cf. Table 2):** Pearson and Spearman correlation coefficients among ground-truth TDA and approximate TDA methods. We show correlations for TDA mean $\hat{\mu}$, TDA standard deviation $\hat{\sigma}$, and TDA p-values.

References

[1] Yann LeCun and Corinna Cortes. The mnist database of handwritten digits. 2005.

[2] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[3] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`, 2022.

[4] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

[5] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020.