Model-agnostic out-of-distribution detection using combined statistical tests

Federico Bergamin[*,1], Pierre-Alexandre Mattei[*,2], Jakob D. Havtorn[1,3], Hugo Senetaire[1], Hugo Schmutz[2,4], Lars Maaløe[1,3], Søren Hauberg[1], Jes Frellsen[1]

Out-of-distribution detection

Consider a curated dataset $x_1, \dots, x_m$ that lives in a space $\mathcal{X}$. Assume we get some new data $\tilde{x}_1, \dots, \tilde{x}_n$. We are interesting in finding if data $\tilde{x}_1, \dots, \tilde{x}_n$ come from:

- the same distribution as $x_1, \dots, x_m$ (IN-DISTRIBUTION)
- a different distribution than $x_1, \dots, x_m$ (OUT-OF-DISTRIBUTION)

Using a one-sided threshold on the log-likelihood of a generative models, as proposed by Bishop (1994), does not work for state-of-the-art deep generative models, as shown by Nalisnick (2018).

We propose a method to **combine** different one-sided test statistics using ***p*-values**, which is **hyperparameter-free** and **works for any differentiable generative model** without relying on model-specific statistics.

Parametric test for OOD detection

Consider a parametric family $(p_\theta)_{\theta \in \Theta}$ of probability densities over $\mathcal{X}$ and learn a suitable $\theta_0 \in \Theta$, for example by fitting a generative model p_{θ_0} on $x_1, \dots, x_m$.

If we assume that $\tilde{x}_1, \dots, \tilde{x}_n \sim_{\text{i.i.d.}} p_{\tilde{\theta}}$ for some unknown $\tilde{\theta} \in \Theta$, we wish to test:

$$\mathcal{H}_0 : \tilde{\theta} = \theta_0,$$
$$\mathcal{H} : \tilde{\theta} \neq \theta_0.$$

The **classic four tests** are:

- likelihood ratio statistic is $S_{LR} = 2(\ell(\hat{\theta}) - \ell(\theta_0))$,
- Wald statistic is $S_W = (\hat{\theta} - \theta_0)^T I(\hat{\theta})(\hat{\theta} - \theta_0)$,
- score statistic is $S_S = \nabla \ell(\theta_0)^T I(\theta_0)^{-1} \nabla \ell(\theta_0)$,
- gradient statistic is $S_G = \nabla \ell(\theta_0)^T (\hat{\theta} - \theta_0)$,

We focus on the **score statistic** for OOD since it is easy to compute, and it does not require fitting an additional model $\hat{\theta}$ on the test points $\tilde{x}_1, \dots, \tilde{x}_n$.

Maximum-mean-discrepancy for OOD detection

Denoting p_{data} the true training data distribution, we can use a *two-sample test*:

$$\mathcal{H}_0 : \tilde{x}_1, \dots, \tilde{x}_n \sim p_{\text{data}},$$
$$\mathcal{H} : \tilde{x}_1, \dots, \tilde{x}_n \not\sim p_{\text{data}}.$$

To measure the distance between p_{data} and $\tilde{x}_1, \dots, \tilde{x}_n$ we need:

- A generative model p_θ to approximate p_{data};
- A measure of distance, we choose **maximum mean discrepancy (MMD)**.

Given a kernel whose feature map is $\Phi : \mathcal{X} \to \mathcal{H}$, the MMD between two distributions P and Q over $\mathcal{X}$ is defined as

$$\text{MMD}_\Phi(P, Q) = \|E_{X \sim P}[\Phi(X)] - E_{Y \sim Q}[\Phi(Y)]\|_{\mathcal{H}}. \tag{1}$$

In our the test statistics will be of the form

$$\text{MMD}_\Phi\left(\frac{1}{m}\sum_{i=1}^m x_i, \frac{1}{n}\sum_{i=1}^n \tilde{x}_i\right) = \left\|\frac{1}{m}\sum_{i=1}^m \Phi(x_i) - \frac{1}{n}\sum_{i=1}^n \Phi(\tilde{x}_i)\right\|_{\mathcal{H}}, \tag{2}$$

Which statistics we should use?

Fisher kernel

$$\Phi_{\text{Fisher}}(x) = I(\theta)^{-\frac{1}{2}} \nabla \log p_\theta(x). \tag{3}$$

$$\text{MMD}_{\Phi_{\text{Fisher}}}\left(\frac{1}{m}\sum_{i=1}^m x_i, \frac{1}{n}\sum_{i=1}^n \tilde{x}_i\right) = \left\|\frac{I(\theta)^{-\frac{1}{2}}}{m}\sum_{i=1}^m \nabla \log p_\theta(x_i) - \frac{I(\theta)^{-\frac{1}{2}}}{n}\sum_{i=1}^n \nabla \log p_\theta(\tilde{x}_i)\right\|_2. \tag{4}$$

- At maximum-likelihood estimate and with no model misspecification: $\mathbb{E}[\nabla \log p_\theta(x)] = 0$.
- The norm of the second term alone then it is equivalent to the **square root of the score statistic**. Due to Occam's Razor, we decide to use the **score statistic** directly.

Typicality kernel

$$\Phi_{\text{Typicality}}(x) = \log p_\theta(x), \tag{5}$$

$$\text{MMD}_{\Phi_{\text{Typicality}}}\left(\frac{1}{m}\sum_{i=1}^m x_i, \frac{1}{n}\sum_{i=1}^n \tilde{x}_i\right) = \left\|\frac{1}{m}\sum_{i=1}^m \log p_\theta(x_i) - \frac{1}{n}\sum_{i=1}^n \log p_\theta(\tilde{x}_i)\right\|_2. \tag{6}$$

- This is equivalent to the **typicality test** proposed by Nalisnick et al, (2019).

Both statistics are **one-sided** and **can be computed for any differentiable generative model**. We show empirically that these are also independent.

Why combining statistics?

Zhang et al. (2021) proved that in case of single-sample OOD detection there is **no test statistic that is constantly better** than all the possible alternatives. We hypothesise that a combination of multiple statistics should perform better, especially in situations where one of the statistics fails.

Proposed method

Our method relies on the computation of the ***p*-values** for the statistics computed on $\tilde{x}_1, \dots, \tilde{x}_n$. We relied on a validation set and **standard bootstrap resampling** procedure to estimate the distribution of the two statistics under $\mathcal{H}_0$.

An optimal way to combine p-values of one-sided independent statistics is the **Fisher's method**:

$$\mathbf{X^2} \sim -\mathbf{2}\sum_{\mathbf{j=1}}^{\mathbf{k}} \ln(\mathbf{p_j}).$$

Technical challenges:

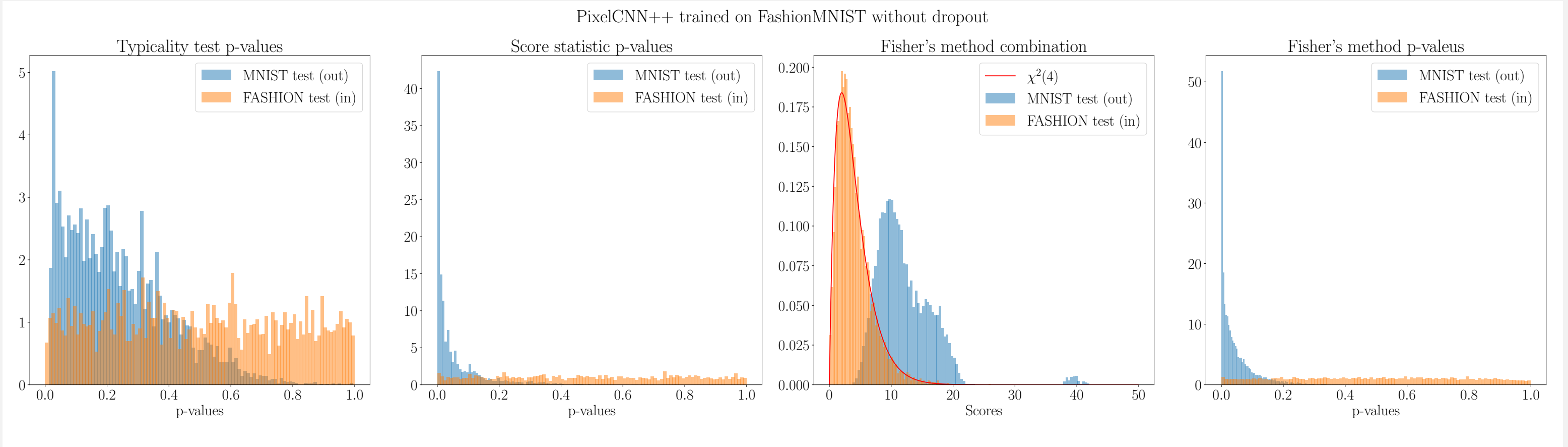
- We use a diagonal approximation for the Fisher Information Matrix $I(\theta)$ using the training set;
- We need to compute per-sample gradients.

References

C. M. Bishop. **Novelty detection and neural network validation** IEE Proceedings-Vision, Image and Signal processing, 1994.
E. Nalisnick et al. **Do deep generative models know what they don't know?**, *ICLR*, 2018.
E. Nalisnick et al. **Detecting out-of-distribution inputs to deep generative models using a test for typicality**, 2019
L. Zhang et al. **Understanding failures in out-of-distribution detection with deep generative models**. *ICML*, 2021
W. Morningstar et al. **Density of states estimation for out of distribution detection**. *AISTATS*. 2021

Are the two statistics really independent?

If the **statistics are independent** and the **null hypotheses are accepted**, then the Fisher combination test statistic χ^2 follows a **chi-squared distribution** with $2k$ degrees of freedom.



Results on single-sample OOD detection

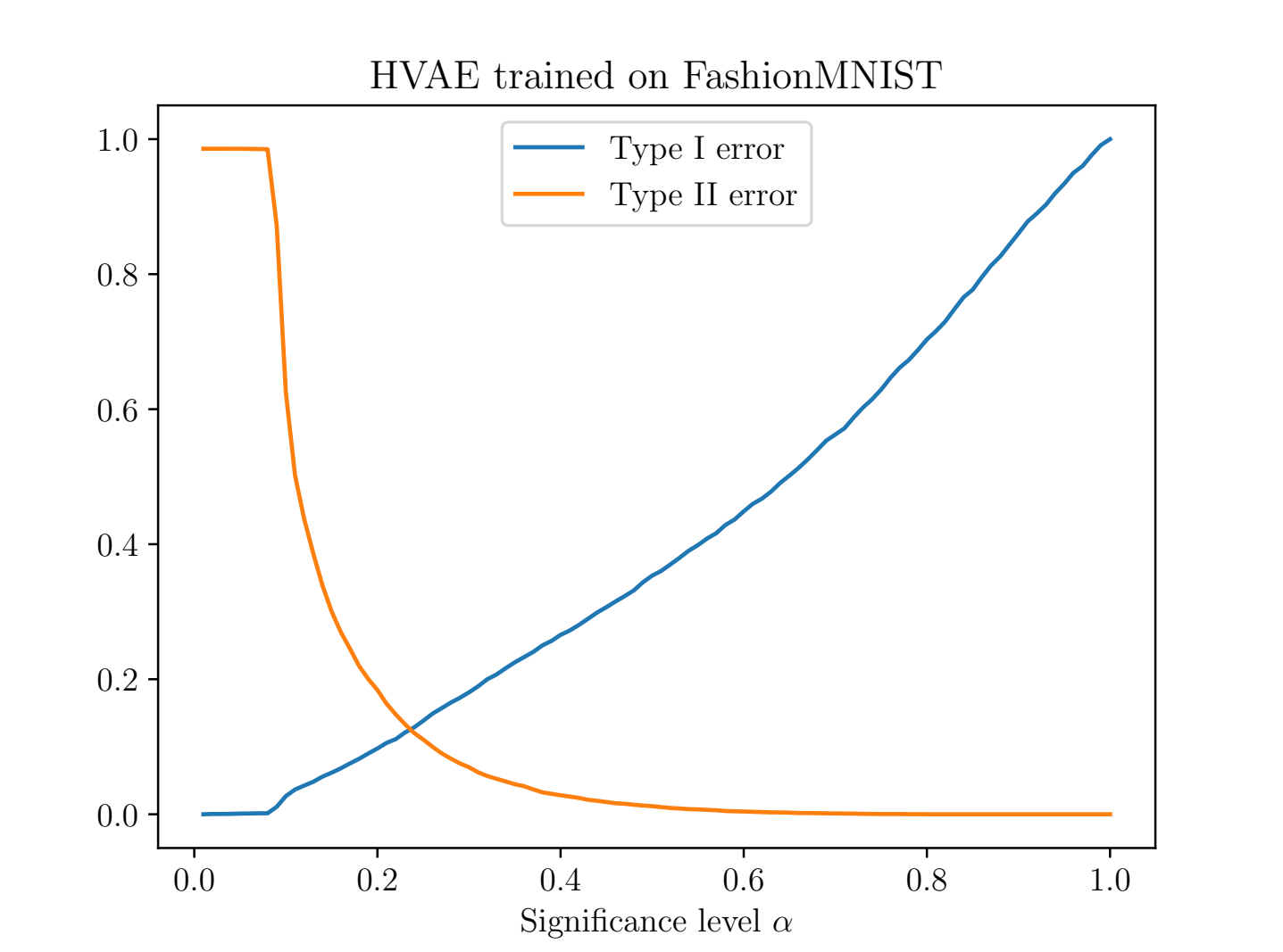
	FashionMNIST (in) / MNIST (out)					
	single statistics				combination	
models	$\log p(x)$	$\|\nabla \log p(x)\|_2$	Typicality	Score Stat	Fisher's method	DoSE$_{\text{KDE}}$
PixelCNN++ (dropout)	0.0762	0.8709	0.8314	**0.8822**	**0.9369**	0.8822
PixelCNN++ (no dropout)	0.1048	**0.9532**	0.7575	0.9381	**0.9536**	0.9382
Glow (RMSProp)	0.1970	0.8904	0.4807	**0.9114**	0.8598	**0.8901**
Glow (Adam)	0.1223	0.7705	0.6987	**0.8745**	**0.8839**	0.8752
HVAE	0.2620	0.8714	0.4884	**0.9578**	0.9383	**0.9498**

	CIFAR10 (in) / SVHN (out)					
	single statistics				combination	
models	$\log p(x)$	$\|\nabla \log p(x)\|_2$	Typicality	Score Stat	Fisher's method	DoSE$_{\text{KDE}}$
PixelCNN++ (model1)	0.1553	**0.8006**	0.6457	0.6407	**0.6826**	0.6571
PixelCNN++ (model2)	0.1567	**0.7923**	0.6498	0.7067	**0.7300**	0.7243
Glow (RMSProp)	0.0630	0.8585	**0.8651**	0.7940	**0.8683**	0.8510
Glow (Adam)	0.0627	0.7844	**0.8624**	0.7655	**0.8613**	0.8588
HVAE	0.0636	0.8067	**0.8679**	0.7335	**0.8603**	0.8179

	CIFAR10 (in) / CIFAR100 (out)					
	single statistics				combination	
models	$\log p(x)$	$\|\nabla \log p(x)\|_2$	Typicality	Score Stat	Fisher's method	DoSE$_{\text{KDE}}$
PixelCNN++ (model1)	0.5153	0.5306	**0.5458**	0.5362	**0.5563**	0.5477
PixelCNN++ (model2)	0.5150	0.5230	**0.5455**	0.5325	**0.5543**	0.5453
Glow (RMSProp)	0.5206	0.5547	0.5507	0.**5801**	**0.5844**	**0.5842**
Glow (Adam)	0.5206	0.5593	0.5508	**0.5692**	**0.5775**	**0.5767**
HVAE	0.5340	0.5280	0.5493	**0.5798**	0.5879	**0.5941**

Can we avoid discarding too many inliers?

When we perform OOD detection, we want to be sure to not discard too many inliers. Using p-values allows us to use well-studied techniques for **false discovery rate (FDR) control**, i.e. controlling the percentage of in-distribution data classified as outliers. We used **Benjamini-Hochberg correction** which guarantees that, for a given significance level α, the FDR stays below that specific level.



[1] Technical University of Denmark, [2] Université Côte d'Azur, Inria, LJAD, CNRS, [3] Corti AI, [4] TIRO, CEA