Evaluating Two-Sample Tests for validating generators in precision sciences

Samuele Grossi[(†) 1,2*], Marco Letizia[2,3*], Riccardo Torre[2*]

[1*] Department of Physics, University of Genova, Via Dodecaneso 33, I-16146 Genova, Italy

[2*] INFN, Sezione di Genova, Via Dodecaneso 33, I-16146 Genova, Italy

[3*] MaLGa-DIBRIS, University of Genova, Via Dodecaneso 35, I-16146 Genova, Italy

† sgrossi@ge.infn.it

INFN — UniGe | DIFI — Università di Genova

1. Motivations and purpose of the work

Model based Monte Carlo

- Computationally demanding
- Reliable synthetic data

ML-based generative models

- Faster simulations
- Lower reliability

Necessity to validate data from generators! This can be done using a **two-sample test**, which checks if two independent samples come from the same probability density function (PDF).

- THEORETICALLY: likelihood-ratio is the most powerful test for simple hypothesis. *Need to know* the PDFs generating the samples.
- PRACTICALLY: Underlying PDFs are usually *unknown* when dealing with real data. Need to use metrics that involve only the data.

Purpose of the work: Establish a rigorous statistical procedure based on robust, simple, and interpretable two-sample tests that can serve both for evaluation and for benchmarking more advanced tests.

2. Test statistics

Test-statistic	Definition
Sliced WD [1]	$t_{\text{SW}} = \frac{1}{K}\sum_{\theta\in\Omega_K}\left(\frac{1}{n}\sum_{i=1}^{n} \mid \underline{x}_i^{\theta} - \underline{x}_i'^{\theta} \mid\right)$
Scaled mean KS	$t_{\overline{\text{KS}}} = \frac{1}{d}\sum_{I=1}^{d}\sqrt{\frac{nm}{n+m}}\,\sup_u \mid F_n^I(u) - G_m^I(u) \mid$
Scaled sliced KS	$t_{\text{SKS}} = \frac{1}{K}\sum_{\theta\in\Omega_K}\sqrt{\frac{nm}{n+m}}\,\sup_u \mid F_n^{\theta}(t) - G_m^{\theta}(t) \mid$
MMD_u^2 [2]	$t_{\text{MMD}} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(x^i,x^j) + \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} k(y^i,y^j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x^i,y^j)$
FGD_∞ [3]	$t_{\text{FGD}} = \lim_{n,m\to\infty}\sum_{I=1}^{d}(\mu_{1,n}^I - \mu_{2,m}^I)^2 + \text{tr}\left(\Sigma_{1,n} + \Sigma_{2,m} - 2\sqrt{\Sigma_{1,n}\Sigma_{2,m}}\right)$
Log-likelihood ratio	$t_{\text{LLR}} = -2\log\frac{\mathcal{L}_{H_0}}{\mathcal{L}_{H_1}}$

3. Reference and Deformed Models

Toy Distributions:

- d dimensional multivariate Correlated Gaussians
- q components, d dimensional mixture of multivariate Gaussians $d = 5, 20, 100$

JetNet Datasets:

- Individual particles in the gluon initiated jets
- Overall jet features

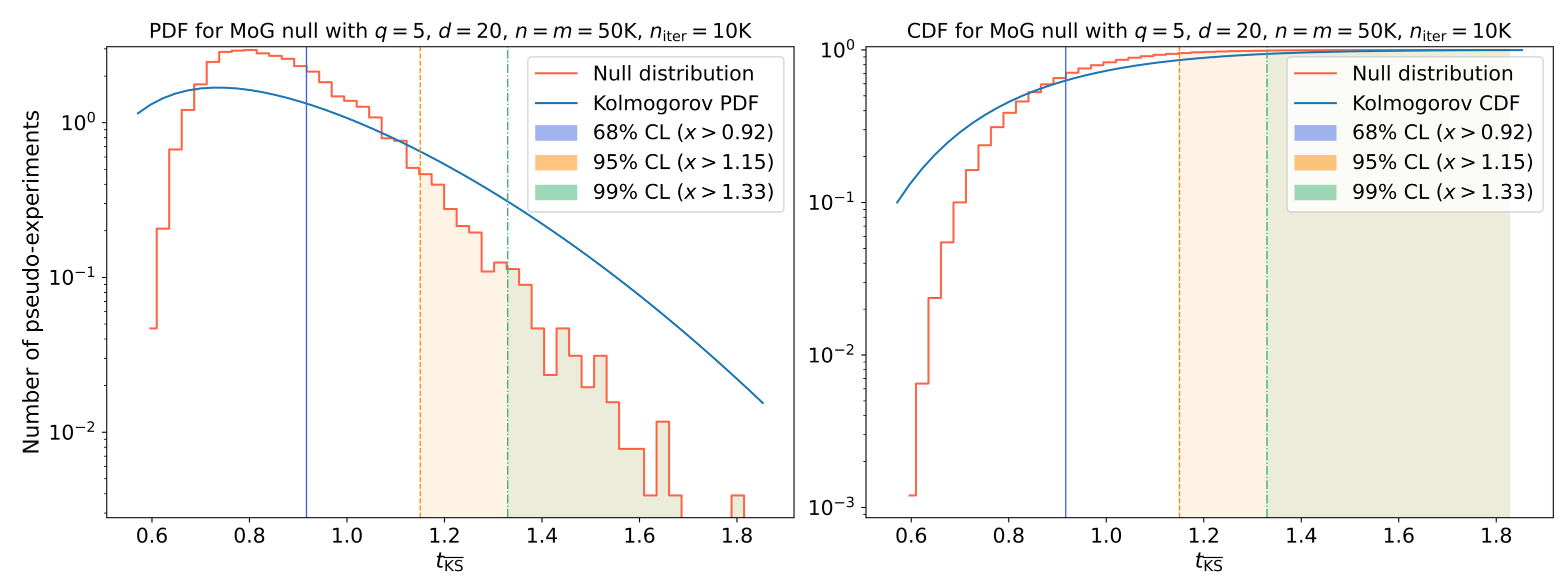
Deformed models are defined by a single parameter ϵ:

(1)	μ-deformation:	$y_{iI} = x_{iI} + \delta_{\mu\, I}$,	$\delta_{\mu\, I} \sim \mathcal{U}_{[-\epsilon,\epsilon]}$
(2)	Σ_{II}-deformation:	$y_{iI} = \mu_I + c_{\Sigma\, I}(x_{iI} - \mu_I)$,	$c_{\Sigma\, I} \sim \mathcal{U}_{[1,1+\epsilon]}$
(3)	$\Sigma_{I\neq J}$-deformation:	$y_{iI} = \sum_j P_{ij}^{(I)} x_{jI}$,	$P_{ij}^{(I)} = P_{ij}^{(I)}(\epsilon)$
(4)	pow_+-deformation:	$y_{iI} = \text{sign}(x_{iI})\lvert x_{iI}\rvert^{1+\epsilon}$,	$\epsilon \geq 0$
(5)	pow_--deformation:	$y_{iI} = \text{sign}(x_{iI})\lvert x_{iI}\rvert^{1-\epsilon}$,	$\epsilon \geq 0$
(6)	$\mathcal{N}$-deformation:	$y_{iI} = x_{iI} + \delta_{iI}$,	$\delta_{iI} \sim \mathcal{N}_{0,\epsilon}$
(7)	$\mathcal{U}$-deformation:	$y_{iI} = x_{iI} + \delta_{iI}$,	$\delta_{iI} \sim \mathcal{U}_{[-\epsilon,\epsilon]}$

4. Methodology and test features

Goal: Make inference on ϵ, finding the smallest value we are sensitive to.

Test H_0: build test statistic distribution under H_0. Perform $10^4(10^3)$ repeated tests on samples drawn from the reference toy distribution(dataset).



Test H_1: perform 100 test on samples extracted from the reference and the deformed distributions. Calculate the mean and standard deviation.

- *test close to the decision boundary*: ϵ such that the mean is at the CL threshold. Use the standard deviation to set an error on ϵ.
- *test different precision*: evaluate each metric varying sample sizes.

5. Example: Results for MoG

MoG model with $\mathbf{d = 20}$, $\mathbf{q = 5}$, and $\mathbf{n = m = 5 \cdot 10^4}$

Statistic	μ-deformation $\epsilon_{95\%\text{CL}}$	$\epsilon_{99\%\text{CL}}$	t (s)	Σ_{ii}-deformation $\epsilon_{95\%\text{CL}}$	$\epsilon_{99\%\text{CL}}$	t (s)	$\Sigma_{i\neq j}$-deformation $\epsilon_{95\%\text{CL}}$	$\epsilon_{99\%\text{CL}}$	t (s)	pow_+-deformation $\epsilon_{95\%\text{CL}}$	$\epsilon_{99\%\text{CL}}$	t (s)
t_{SW}	$0.04957^{+0.018}_{-0.02}$	$0.06694^{+0.017}_{-0.017}$	3023	$0.01679^{+0.005}_{-0.0063}$	$0.02315^{+0.0045}_{-0.005}$	3197	$0.02162^{+0.0056}_{-0.008}$	$0.02935^{+0.0045}_{-0.0055}$	**3410**	$0.00581^{+0.0017}_{-0.0022}$	$0.00798^{+0.0015}_{-0.0017}$	**3157**
$t_{\overline{\text{KS}}}$	$\mathbf{0.00482^{+0.0013}_{-0.0018}}$	$\mathbf{0.00667^{+0.0011}_{-0.0013}}$	2966	$\mathbf{0.00175^{+0.00052}_{-0.00068}}$	$\mathbf{0.00248^{+0.00042}_{-0.00052}}$	3185	$1.00146^{+0.00074}_{-0.00031}$	$1.00238^{+0.00055}_{-0.00031}$	3967	$\mathbf{0.0004^{+0.00015}_{-0.00017}}$	$\mathbf{0.00059^{+0.00013}_{-0.00014}}$	3363
t_{SKS}	$0.03647^{+0.011}_{-0.014}$	$0.04821^{+0.011}_{-0.012}$	**2899**	$0.01329^{+0.003}_{-0.0043}$	$0.01759^{+0.0025}_{-0.003}$	**3022**	$0.02306^{+0.0071}_{-0.0088}$	$0.03079^{+0.0062}_{-0.0072}$	3553	$0.0043^{+0.0009}_{-0.0013}$	$0.00565^{+0.00074}_{-0.0009}$	3193
t_{FGD}	$0.05778^{+0.026}_{-0.027}$	$0.0787^{+0.023}_{-0.021}$	4047	$0.01945^{+0.0063}_{-0.0081}$	$0.02651^{+0.0053}_{-0.0056}$	4507	$\mathbf{0.00551^{+0.0015}_{-0.002}}$	$\mathbf{0.00748^{+0.0013}_{-0.0013}}$	6327	$0.00702^{+0.0021}_{-0.0028}$	$0.00965^{+0.0016}_{-0.0019}$	4870
t_{MMD}	$0.04425^{+0.019}_{-0.018}$	$0.06215^{+0.017}_{-0.015}$	10204	$0.00923^{+0.0058}_{-0.0051}$	$0.01305^{+0.0053}_{-0.0044}$	11217	$0.01723^{+0.008}_{-0.0072}$	$0.02431^{+0.0069}_{-0.0064}$	11450	$0.00332^{+0.0018}_{-0.0017}$	$0.00467^{+0.0017}_{-0.0014}$	11801
t_{LLR}	$0.00021^{+0.00013}_{-0.00014}$	$0.0003^{+0.00013}_{-0.00014}$	5911	$0.00007^{+0.00005}_{-0.00004}$	$0.0001^{0.00005}_{-0.00004}$	6304	-	-	-	$0.00002^{+0.00001}_{-0.00001}$	$0.00002^{+0.00001}_{-0.00001}$	6877

Statistic	pow_--deformation $\epsilon_{95\%\text{CL}}$	$\epsilon_{99\%\text{CL}}$	t (s)	$\mathcal{N}$-deformation $\epsilon_{95\%\text{CL}}$	$\epsilon_{99\%\text{CL}}$	t (s)	$\mathcal{U}$-deformation $\epsilon_{95\%\text{CL}}$	$\epsilon_{99\%\text{CL}}$	t (s)	Timing t^{null} (s)
t_{SW}	$0.00604^{+0.0017}_{-0.0023}$	$0.00825^{+0.0016}_{-0.0018}$	**3051**	$0.19318^{+0.025}_{-0.039}$	$0.22704^{+0.019}_{-0.026}$	**2403**	$0.33394^{+0.044}_{-0.068}$	$0.39248^{+0.033}_{-0.044}$	**2354**	338
$t_{\overline{\text{KS}}}$	$\mathbf{0.00042^{+0.00015}_{-0.00018}}$	$\mathbf{0.00061^{+0.00013}_{-0.00015}}$	3372	$\mathbf{0.00751^{+0.002}_{-0.0024}}$	$\mathbf{0.00993^{+0.0018}_{-0.002}}$	2934	$\mathbf{0.01211^{+0.003}_{-0.0035}}$	$\mathbf{0.01575^{+0.0027}_{-0.003}}$	2835	**155**
t_{SKS}	$0.00441^{+0.00092}_{-0.0014}$	$0.00574^{+0.00077}_{-0.00094}$	3324	$0.15874^{+0.023}_{-0.034}$	$0.18473^{+0.019}_{-0.023}$	2726	$0.27395^{+0.041}_{-0.059}$	$0.3188^{+0.033}_{-0.04}$	2601	509
t_{FGD}	$0.00722^{+0.0021}_{-0.0027}$	$0.00987^{+0.0016}_{-0.0019}$	4892	$0.18095^{+0.023}_{-0.038}$	$0.21269^{+0.016}_{-0.02}$	3756	$0.31409^{+0.04}_{-0.07}$	$0.36919^{+0.027}_{-0.036}$	3643	2795
t_{MMD}	$0.00353^{+0.0016}_{-0.0015}$	$0.00494^{+0.0014}_{-0.0012}$	11418	$0.43531^{+0.066}_{-0.11}$	$0.51609^{+0.045}_{-0.054}$	8642	$0.75353^{+0.12}_{-0.18}$	$0.89336^{+0.078}_{-0.098}$	7700	13860
t_{LLR}	$0.00002^{+0.00001}_{-0.00001}$	$0.00002^{+0.00001}_{-0.00001}$	6991	-	-	-	-	-	-	-

6. Conclusions

- The likelihood ratio, when calculable, shows about an order of magnitude greater sensitivity compared to the other metrics.
- The metrics based on 1D tests (t_{SW}, $t_{\overline{\text{KS}}}$, t_{SKS}) are easy to implement regardless of sample sizes and scale linearly with dimensions. In contrast, FGD_∞ requires large sample sizes to perform well, while MMD_u^2 suffers the curse of dimensionality in such cases. Furthermore, despite their simplicity, 1D-based metrics show high sensitivity to all deformations, being therefore applicable to a wide range of scenarios.
- Similar sensitivity across different datasets for the same deformation, proves that our procedure provides a robust evaluation of the tests themselves.
- We think the proposed test statistics could serve as a valuable first step in evaluating a generator, before considering more resource-intensive tools.

References

[1] N. Bonneel, J. Rabin, G. Peyré and H. Pfister, *"Sliced and Radon Wasserstein Barycenters of Measures"*. In: Journal of Mathematical Imaging and Vision 51 (2015) 22

[2] A. Gretton, K. M. Borgwardt, M. J. Rasch., B. Schölkopf and A. Smola. *"A Kernel Two-Sample Test"*. In: Journal of Machine Learning Research 13 (2012) 723-773.

[3] R. Kansal, A. Li, J. Duarte, N. Chernyavskaya, M. Pierini, B. Orzari and T. Tomei. *"Evaluating generative models in high energy physics"*. In: Phys. Rev. D (2023).