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

$t_{\mathrm{SW}} = \frac{1}{K}\sum_{\theta\in\Omega_K}\left(\frac{1}{n}\sum_{i=1}^{n} \mid \underline{x}_i^{\theta} - \underline{x}_i'^{\theta} \mid\right)$

$t_{\overline{\mathrm{KS}}} = \frac{1}{d}\sum_{I=1}^{d}\sqrt{\frac{nm}{n+m}}\,\sup_u \mid F_n^I(u) - G_m^I(u)|$

$t_{\mathrm{SKS}} = \frac{1}{K}\sum_{\theta\in\Omega_K}\sqrt{\frac{nm}{n+m}}\,\sup_u \mid F_n^{\theta}(t) - G_m^{\theta}(t) \mid$

$t_{\mathrm{MMD}} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(x^i,x^j) + \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} k(y^i,y^j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x^i,y^j)$

$t_{\mathrm{FGD}} = \lim_{n,m\to\infty}\sum_{I=1}^{d}(\mu_{1,n}^I - \mu_{2,m}^I)^2 + \mathrm{tr}\left(\Sigma_{1,n} + \Sigma_{2,m} - 2\sqrt{\Sigma_{1,n}\Sigma_{2,m}}\right)$ [1]

$t_{\mathrm{NPLM}} = -2\left[\frac{n}{m}\sum_{x\in\mathcal{X}}\left(e^{f_{\hat{w}}(x)} - 1\right) - \sum_{y\in\mathcal{Y}} f_{\hat{w}}(y)\right], \qquad f_w = \sum_{i=1}^{M} w_i k_{\sigma}(\tilde{x},\tilde{x}_i)$ [2]

$t_{\mathrm{LLR}} = -2\log\frac{\mathcal{L}_{H_0}}{\mathcal{L}_{H_1}}$

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:

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

Features number = 3, 90

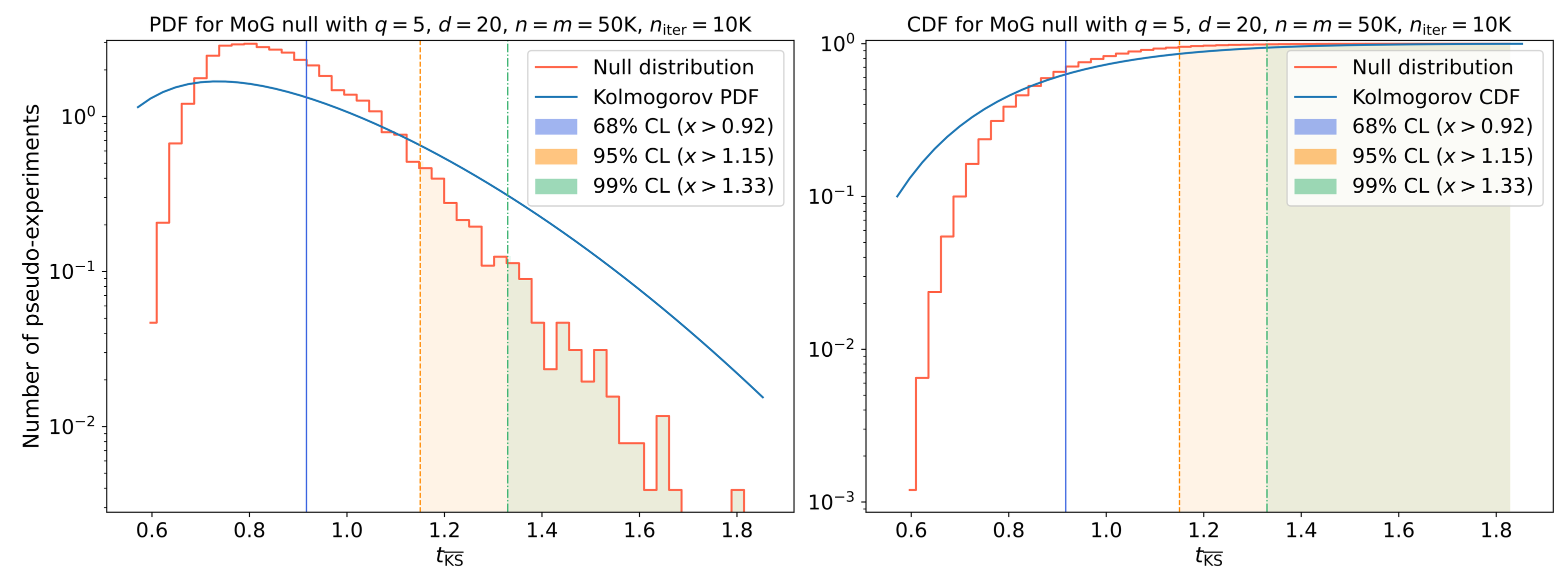
Deformed models are defined by a single parameter ϵ:

(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:	$y_{iI} = \mathrm{sign}(x_{iI})\lvert x_{iI}\rvert^{1+\epsilon}\,,$	$\epsilon \geq 0$
(5)	pow$_-$-deformation:	$y_{iI} = \mathrm{sign}(x_{iI})\lvert x_{iI}\rvert^{1-\epsilon}\,,$	$\epsilon \geq 0$
(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.

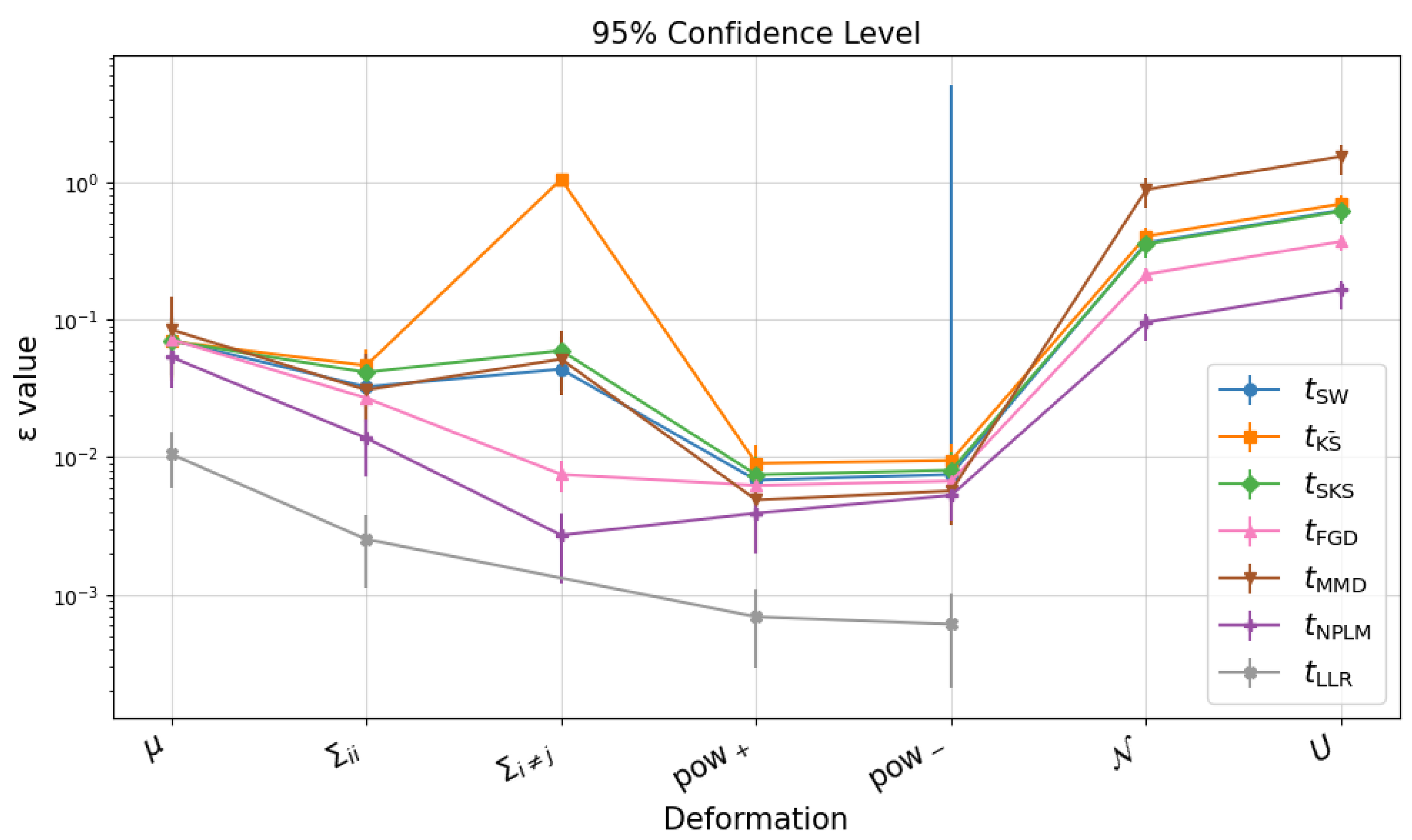
Test H_0: build test statistic distribution under H_0. Perform $10^4(10^3)$ repeated tests on samples drawn from the reference toy distribution(dataset).

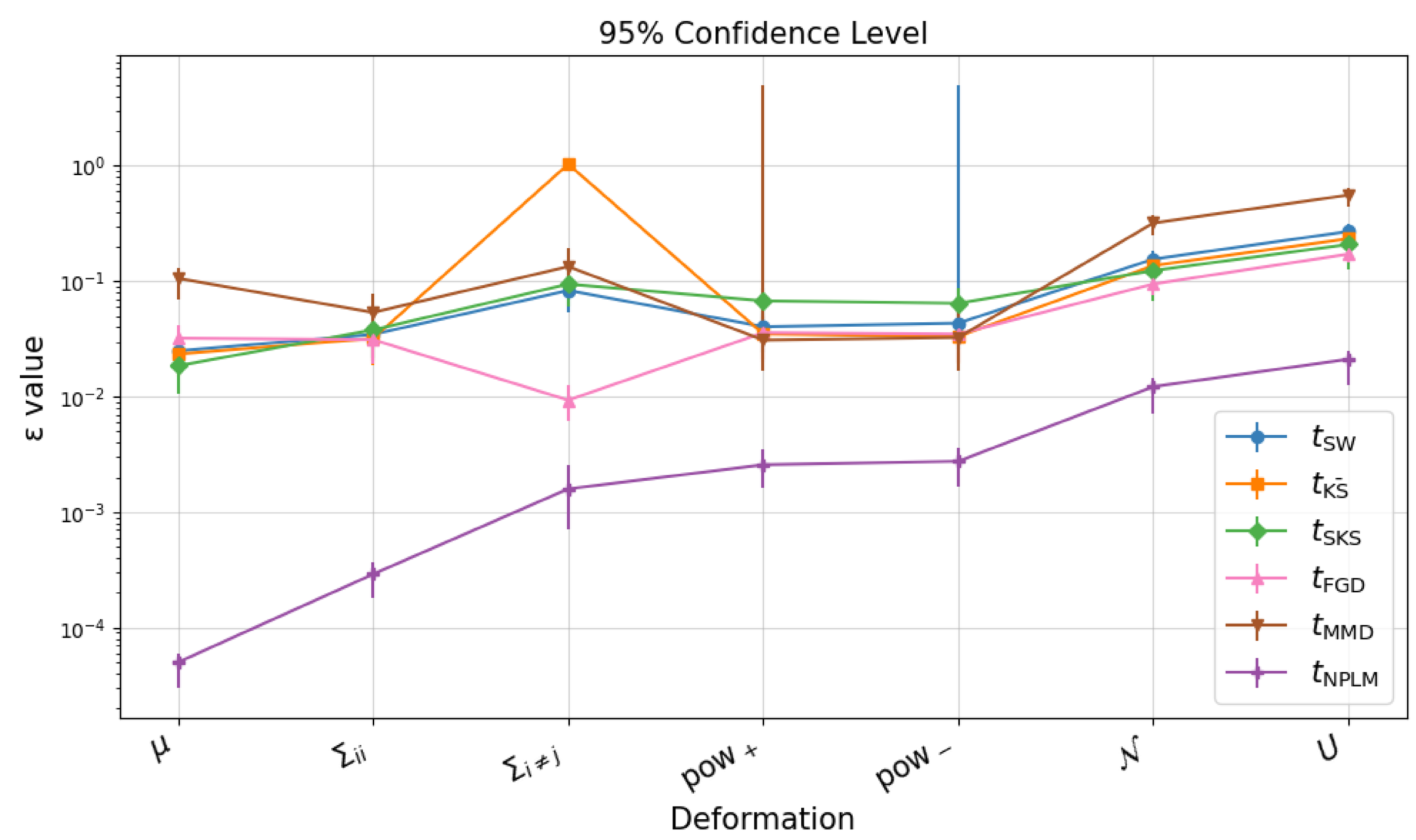


Test H_1: perform 100 test on samples extracted from the reference and the deformed distributions. Calculate the mean and standard deviation to get a central value and an error on ϵ

- *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. Some Results (MoG and Jet features)





6. Conclusions

- 1D based metrics are robust and efficient, while NPLM is more sensitive but slower.
- No universally best metric. When speed is prioritized over precision, 1D based metrics are preferable. When sensitivity is the main goal and time is not a limiting factor, NPLM is more effective.
- A good strategy is to start with 1D based metrics for a quick validation. If they do not reject the model, switch to more ML metrics, like NPLM.
- Another possibility is to use the 1D based metrics during model parameters tuning and rely on more powerful metrics for the validation.

References

[1] 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).

[2] G. Grosso, M. Letizia. *"Multiple testing for signal-agnostic searches of new physics with machine learning"*. In: European Physics Journal C (2025).

[3] S. Grossi, M. Letizia, R. Torre. URL: https://github.com/TwoSampleTests