Introduction to Statistical Learning in PDF Reconstruction

Emma Allen
Supervisor: Prof. Luigi Del Debbio

Motivation

Accurate Parton Distribution Functions (PDFs) are essential for predicting outcomes in hadron collider experiments, but their non-perturbative nature means they must be constructed from data.

We focus on data from NNPDF, which provides Monte Carlo replicas representing the underlying probability distribution of PDFs. Although NNPDF does not explicitly assume Gaussian behaviour, the replicas often appear Gaussian, motivating an analysis of their statistical properties.

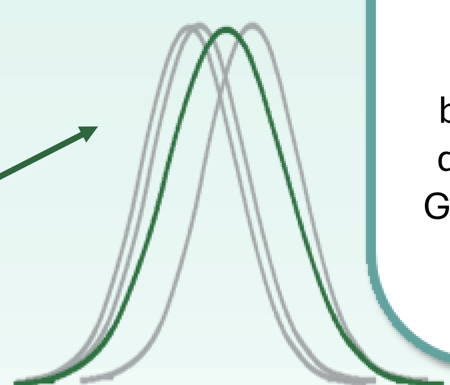
This work has three primary objectives:

1. Assess the Gaussianity of our NNPDF4.0 dataset.
2. Reconstruct PDFs in 2D using Kernel Density Estimation.
3. Construct a global correlation matrix using 2D PDF reconstruction and compare it with its empirical counterpart.

This contributes to a broader effort to understand how assuming Gaussian behaviour in NNPDF data affects the calculation of physical observables derived from these PDFs.

Kernel Density Estimation (KDE)

- KDE estimates the probability density by placing a Gaussian kernel at each data point.
- Adding and normalising these kernels produces a smooth continuous PDF.
- Illustrated by stacking grey Gaussians to form the green curve.

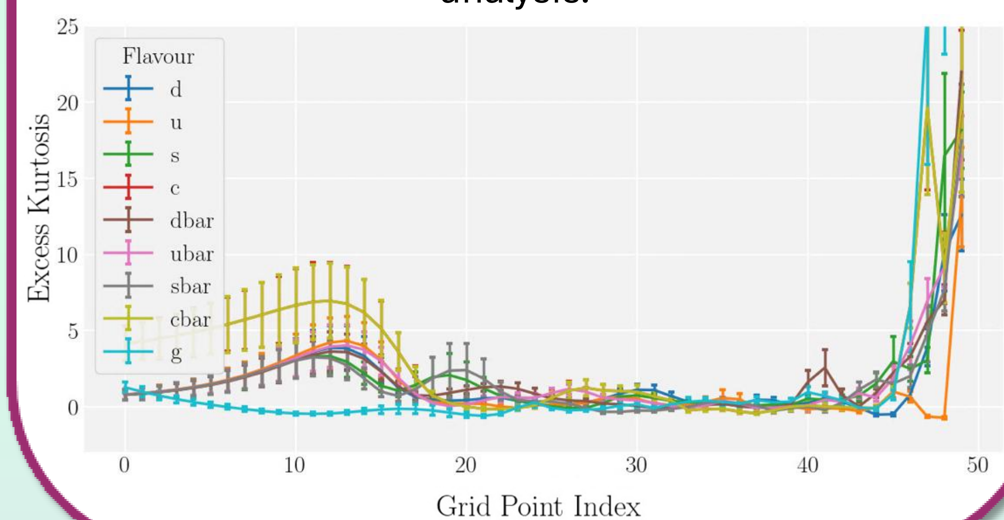


Assessing Gaussianity of the Data

The dataset comprises 1000 replicas evaluated at 50 grid points for nine parton flavours ($u, d, s, c, \overline{u}, \overline{d}, \overline{s}, \overline{c}$, g), yielding a 450-dimensional grid.

Gaussianity is tested via excess kurtosis of each 1D marginal (flavour–grid index pair), with zero indicating a perfect Gaussian.

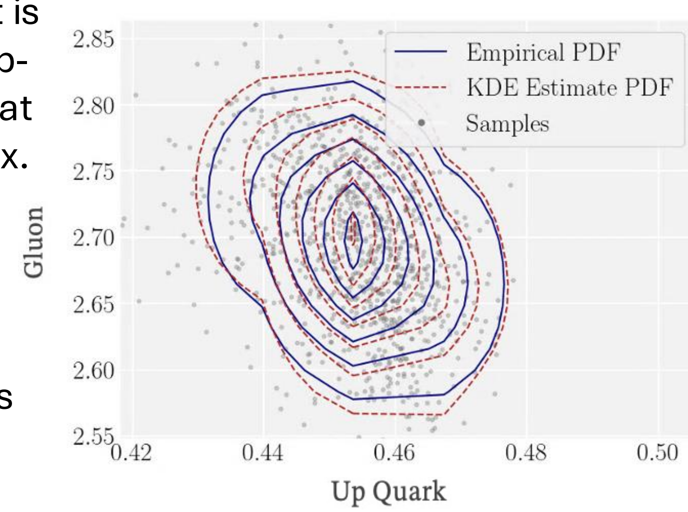
The kurtosis-by-flavour plot below shows approximate Gaussianity up to the 45th grid point where excess kurtosis increases rapidly; only these are used in further analysis.



2D PDF Reconstruction

Shown to the right is the KDE for the up-quark–gluon pair at the 28th grid index.

The up-quark bulging reflects deviations from Gaussianity in its underlying distribution.



Matrix Reconstruction Method

Direct construction of the full 450-dimensional correlation matrix is computationally infeasible. Instead, we

- Compute the covariance for all flavour-grid pairs
- Assemble the full matrix by averaging overlapping sub-blocks
- Normalise the full matrix to get correlation matrix

When computing moments, we have two approaches: KDE with integration (LHS), and empirical averaging (RHS).

Function corresponding to desired moment (e.g. covariance)

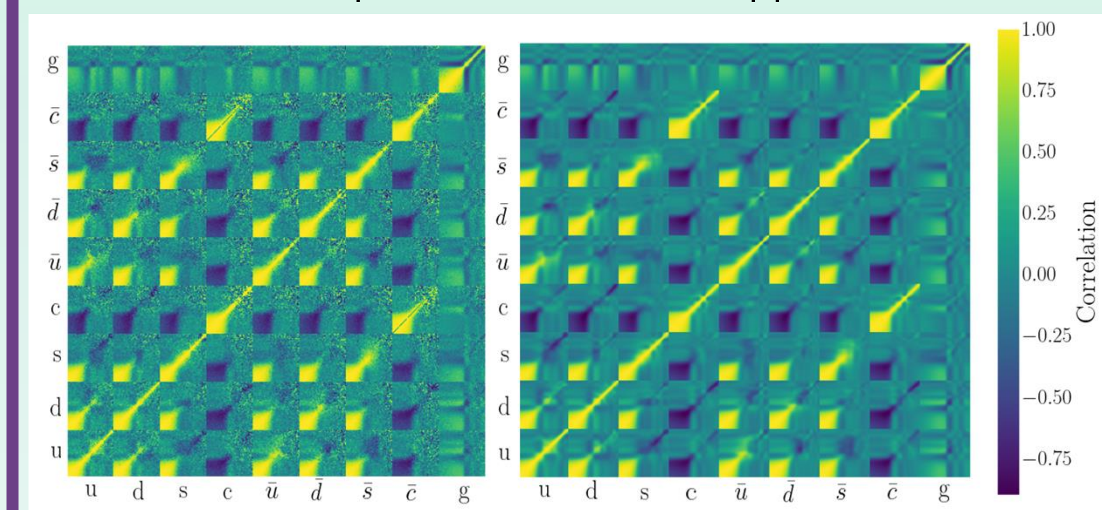
Sample size

$$\int f(\mathbf{X})\hat{p}_{KDE}(\mathbf{X})\,d\mathbf{X} \approx \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}_i)$$

KDE estimator

Correlation Matrix Reconstruction Results

In the figure below, the KDE (left) and empirical (right) matrices show consistent main patterns. The KDE grainy outside these patterns, especially at higher grid indices, likely due to assumptions and numerical approximations.



While sum rules and physical constraints may contribute to these patterns, their exact origin remains unclear.

Future work could explore this and look how error from assuming Gaussianity propagates into calculations of observable quantities.