Introduction to Statistical Learning in PDF Reconstruction

Emma Allen
Supervisor: Prof. Luigi Del Debbio

Motivation

The parton model, first devised in 1969 describes hadrons as a collection of partons, later discovered to be quarks, anti-quarks, and gluons.

Parton Distribution Functions (PDFs) give the probability of finding a parton with a certain momentum fraction at a fixed scale.

Since PDFs cannot be calculated directly, they must be extracted from experimental data and are essential for collider predictions.

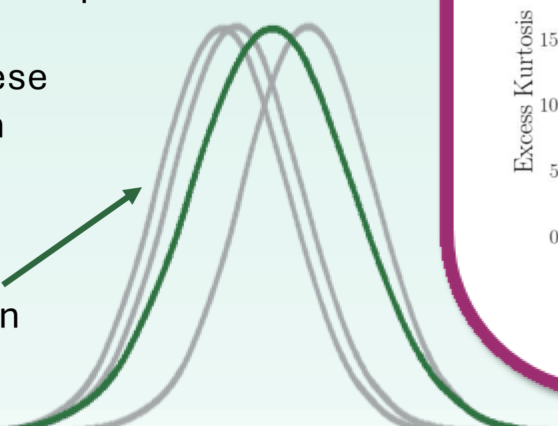
This work has three primary objectives:

1. Assess the Gaussianity of the one-dimensional replicas in NNPDF4.0 data.
2. Reconstruct the PDF in 1D and 2D using Kernel Density Estimation.
3. Using the 2D PDF reconstruction, construct global correlation matrix and compare with its empirical counterpart.

This contributes to a broader effort to understand how assuming Gaussian behaviour in NNPDF4.0 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.

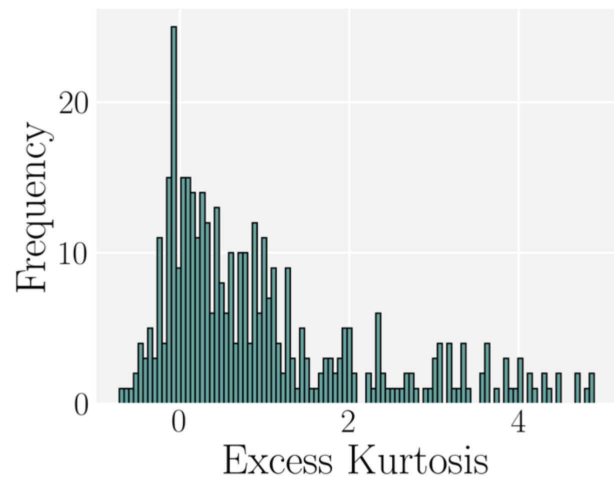


Assessing Gaussianity of the Data

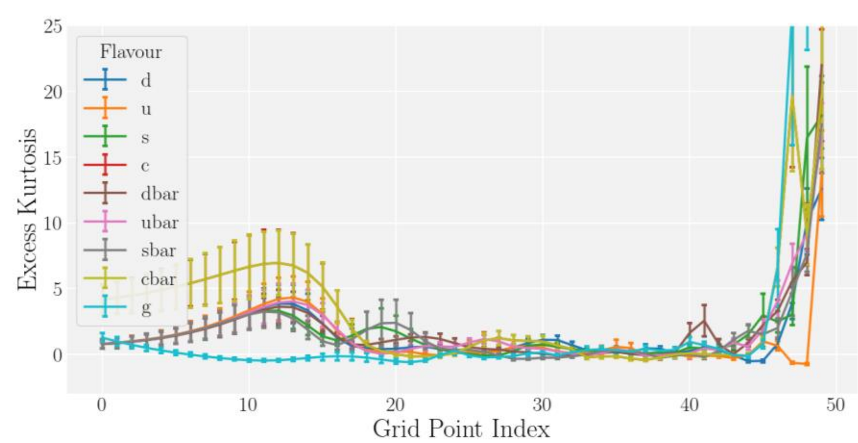
The dataset consists of 1000 replicas (a PDF) evaluated at 50 grid points for each of nine chosen parton flavours ($u, d, s, c, \bar{u}, \bar{d}, \bar{s}, \bar{c}$, g), This results in a 450-dimensional grid of replicas.

Gaussianity is tested by decomposing the dataset into 450 one-dimensional marginals, one for each flavour–grid point pair. The excess kurtosis (fourth standardised moment) is computed for each marginal, where zero indicates a perfect Gaussian.

The histogram below shows a strong peak near zero, indicating that most marginals follow a Gaussian



The kurtosis by flavour plot beneath shows the marginals remain approximately Gaussian up to the 45th grid point. Therefore, only the first 45 grid points are used in matrix reconstructions.



Global Matrix Reconstruction Method

Direct construction of the full 450-dimensional covariance matrix is computationally infeasible.

Instead, we compute 2D covariance estimates for variable pairs and assemble the full matrix by averaging overlapping sub-blocks. Moments are then computed with two approaches: KDE with integration (LHS) , and empirical averaging (RHS).

KDE estimator

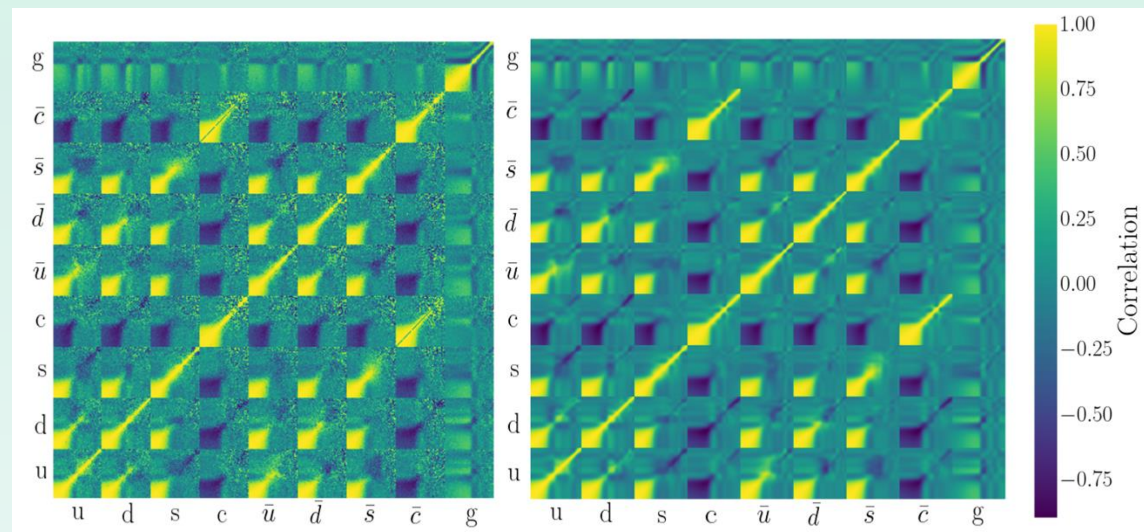
Sample size

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

Function corresponding to desired statistic (e.g. $f(x) = x$ for the mean or $f(x, y) = (x-\mu_x)(y-\mu_y)$ for the covariance, which is normalised to give the correlation)

Correlation Matrix Reconstruction Results

Both the KDE (left) and empirical (right) show the same overall pattern, though the KDE looks grainy at higher indices, likely from Gaussian assumptions and numerical approximations.



Sum rules and physical constraints may play a role in explaining this pattern, but the exact origin remains unclear and requires further study.