Introduction to Statistical Learning in Parton Distribution Function Reconstruction

Emma Allen

Supervisor: Prof. Luigi Del Debbio

August 2025

Abstract

Parton Distribution Functions (PDFs) are essential for predicting outcomes in hadron collider experiments, however their non-perturbative nature prevents direct analytical calculation and must be determined from experimental data. This paper has three primary objectives. First, we assess the Gaussianity of the one-dimensional replicas in NNPDF4.0 by computing their excess kurtosis, finding that the majority of replicas are approximately Gaussian. Second, we reconstruct the PDF in both one and two dimensions using Kernel Density Estimation (KDE), from these reconstructions, statistical moments are computed and found to be consistent with their empirically calculated counterparts within their associated uncertainties. Finally, the two dimensional covariance estimates are then combined and normalised to produce a global correlation matrix, which shows strong visual agreement with the empirical reconstruction. This work 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.

Contents

1 Introduction

Precise knowledge of the internal structure of hadrons is crucial for accurately predicting outcomes in hadron collider experiments and, more broadly for searching for new physics in high-energy phenomenology. PDFs are central to this understanding, as they describe how the momentum of a nucleon is shared among its constituent partons (quarks, anti-quarks and gluons). For instance at the Large Hadron Collider (LHC), accurate determination of PDFs has proved essential for a wide range of studies from Higgs boson characterisation and precision Standard Model measurements to searches for new physics [1, 2].

Since the start of LHC Run I, methods for determining PDFs have been progressively refined and expanded to incorporate new data. Because PDFs are inherently non-perturbative, they cannot be calculated from first principles; instead, they are determined by fitting a wide range of experimental data from multiple experiments. Leading collaborations combine measurements from experiments such as Deep Inelastic Scattering (DIS), Drell–Yan production, and jet production. Notable examples include CTEQ [3, 4] and MMHT [5] which use Hessian-based methods to express PDFs via predetermined functional forms with a fixed number of parameters. In these cases, uncertainties are modelled by assuming a multivariate Gaussian distribution in parameter space [6]. Another major collaboration is NNPDF [7–11], which models PDFs using neural networks trained on Monte Carlo replicas of the experimental data. Each replica corresponds to a resampled dataset that incorporates experimental uncertainties. This produces an ensemble of PDFs, providing a Monte Carlo representation of the underlying probability distribution. Uncertainties are then obtained directly from statistical moments of this ensemble, without assuming any fixed functional form or shape [6]. All of these global fits have steadily improved by incorporating more recent datasets, notably from LHC Run II.

For our analysis, we focus on the NNPDF collaboration. This collaboration began with NNPDF1.0 [7] by using neural networks trained on Monte Carlo replicas to model PDFs. Subsequent versions improved and expanded this framework: NNPDF2.0 [9] enlarged the dataset and refined neural network training; NNPDF3.0 [8] incorporated extensive LHC Run I data and improved the treatment of theoretical uncertainties; and the latest NNPDF4.0 [11] introduced enhanced neural architectures and included LHC Run II data. More recently, the SIMUnet methodology by Iranipour and Ubiali [12] extended this framework by adding a layer to simultaneously determine PDFs alongside an arbitrary number of related parameters.

Although the NNPDF methodology does not make explicit Gaussian assumptions, replica ensembles often exhibit Gaussian-like behaviour. Motivated by this observation, this work addresses two main goals: first, to understand the Gaussian characteristics of the NNPDF replica data itself. Second, to perform one and two dimensional PDF reconstructions using Kernel Density Estimation (KDE). Overall this work contributes to a broader effort to understand how assuming Gaussian behaviour in NNPDF data affects the calculation of physical observables derived from these PDFs.

This paper is organised as follows: sections two and three summarise the key principles of DIS and present the mathematical framework required for this study. Section four details our methodology, dataset and assumptions for performing one and two dimensional KDE PDF reconstructions, which are then used to construct global covariance and correlation matrices. Finally section five presents and analyses the results we obtain.

2 Background: Overview of Deep Inelastic Scattering (DIS)

In particle physics, some of the most important experiments for probing the internal structure of nucleons are Deep Inelastic Scattering (DIS) experiments, in which high-energy leptons scatter off nucleons. These experiments were first developed in the 1960s and 1970s, they were instrumental in establishing the quark model of hadronic matter [13]. The theoretical framework for interpreting DIS is built upon two key concepts: the parton model and Bjorken scaling.

The parton model, proposed by Richard Feynman, is essentially a formal statement of the notion that the nucleon is made up of smaller constituents, partons [14]. No initial assumptions about partons were made, however later experiments and developments in Quantum Chromodynamics (QCD) revealed them to be quarks, antiquarks and gluons. The PDF, $f_i(x)$ gives the probability of finding a parton of type i carrying a fraction x of the nucleon's momentum. In other words they describe the momentum distribution among the nucleon's constituent partons. These functions are crucial for making predictions in hadron collider physics and calculating observables (e.g. structure functions and cross sections). However, as noted earlier, PDFs cannot be derived directly from theory because QCD becomes strongly coupled at low energies [13]; therefore, they must be inferred from experimental data.

The second key idea is Bjorken Scaling, predicted by James Bjorken. Stated simply, it is the prediction that when the momentum carried by the probe becomes very large, then the cross section's dependence on parameters like energy and momentum squared becomes very simple. Or in other words, at high momentum transfer, Bjorken scaling implies that the interaction between a high-energy probe and a nucleon can be approximated as a direct, one to one scattering between a single probe particle (e.g. an electron) and a single parton inside the nucleon (e.g. an up quark). This insight simplifies the parton model considerably and allows us to write PDFs solely in terms of the Bjorken variable, which is interpreted as the fraction of the nucleon's momentum carried by the parton, x as previously defined [13]. We observe this behaviour when the wavelength of the probe is much less than the nucleon diameter. This implies a probe momentum above roughly 1 GeV [13].

In summary, PDFs characterise how the nucleon's momentum is distributed among its constituent partons. These functions are essential for calculating observables in hadron collider experiments and the concept of Bjorken Scaling significantly simplifies them. Improved accuracy in PDF reconstruction directly enhances the precision of such calculations, making their reconstruction a crucial task. We now present the mathematical framework underlying our reconstruction method.

3 Background: Mathematical Foundations

The mathematical basis of this project is two fold. Firstly, we use non-parametric estimation; a set of statistical methods that infer the underlying probability distribution directly from the data, without assuming a fixed parametric form [15]. Second, we use statistical moments to quantify the shape of our distributions. We begin by introducing two non-parametric methods: histograms and Kernel Density Estimation (KDE), and applying them to synthetic Gaussian datasets as an initial demonstration. We then explain statistical moments and show how both techniques are used together in our analysis.

3.1 Histogram Estimation Technique

The histogram method estimates the probability density by partitioning the input space into bins and counting the number of data points in each bin, without assuming any specific distribution [15]. To illustrate this approach using a standard Gaussian distribution, we plot the histogram estimate alongside its corresponding analytical Probability Density Function (PDF), defined below (throughout the rest of this paper, PDF will refer to Probability Density Function)

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

where x is a random variable, μ is the expectation (mean) and σ is the standard deviation. The corresponding histogram for a standard Gaussian is presented in Figure 1.

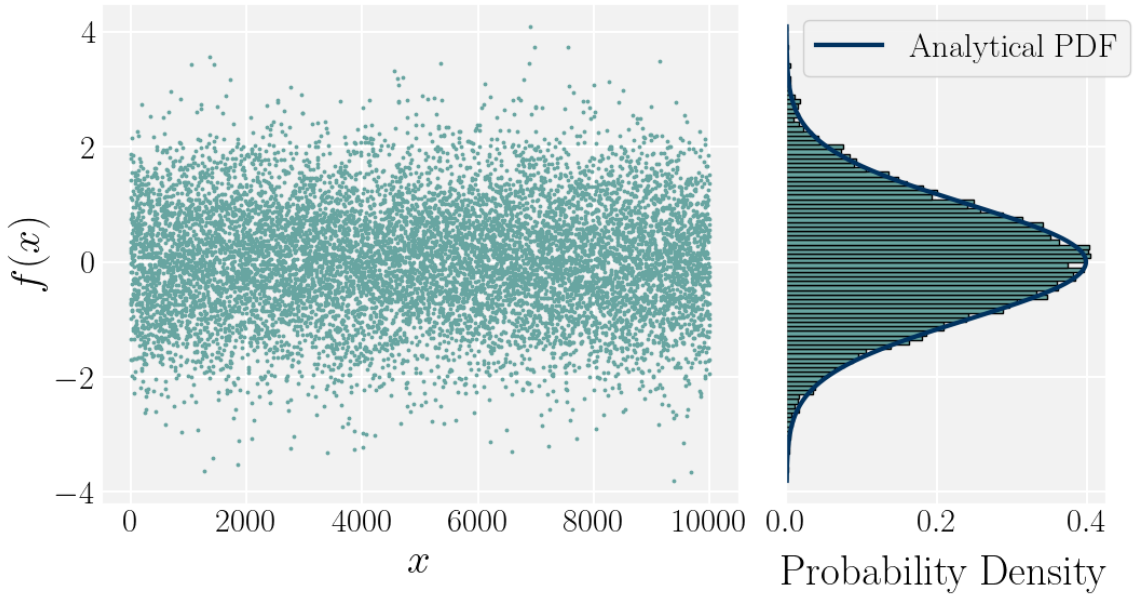

Figure 1: Scatter plot of a standard Gaussian distribution (left) alongside its histogram (right) with the analytical PDF (solid blue) superimposed. Parameters: $\mu = 0$, $\sigma = 1$, sample size $n = 10000$.

With a large enough sample size ($n \gtrsim 1000$), the histogram estimation performs well; however, since the probability density is assumed constant within each bin, discontinuities arise at the bin boundaries. This is not useful in practise where we require a continuous estimate. Thus we must turn to alternative methods that give a continuous estimate, such as KDE.

3.2 Kernel Density Estimation (KDE)

KDE is a non-parametric method for estimating the PDF of a random variable [15]. It constructs the estimate by centring a smooth kernel function, typically Gaussian, at each data point. Then summing the contributions of all kernels to form a continuous density function. It can be understood as constructing a superposition of Gaussian kernels that share a common standard deviation (determined by the bandwidth parameter) but have a different mean corresponding to each data point, with the entire sum normalised by a scaling factor. In KDE, it is helpful to distinguish between univariate and multivariate cases.

In the univariate case, the kernel's shape and smoothness are controlled by a single bandwidth parameter h. This parameter determines how much the density estimate is smoothed. A small

bandwidth parameter closely follows the data but can be noisy, while a large bandwidth parameter produces a smoother estimate but can hide important details. There is no single method for calculating the optimal bandwidth. Different methods exist which include analytical and data-driven approaches. Here, we use Cross Validation (CV) to choose the bandwidth parameter.

In this method, the dataset is divided into five equal parts (folds). For each candidate bandwidth, the KDE is constructed using four folds, while the remaining fold serves as the test set. As the name suggests, the test fold is used to assess how well the estimate fits new data by calculating the log-likelihood of the test fold points. We then rotate which fold is the test fold and repeat the process so that each fold is used once for testing. The average log-likelihood across all five folds is computed for each candidate bandwidth. The bandwidth that maximises the average log-likelihood is selected as optimal and denoted h_{opt}. The final KDE is computed as

$$\hat{p}_{\mathrm{KDE}}(x) = \frac{1}{nh_{opt}} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_{opt}}\right),$$

where n is the number of data points, K is the kernel function (we assume Gaussian) and x_i are the data points. With a Gaussian kernel, this expression becomes

$$\hat{p}_{\mathrm{KDE}}(x) = \frac{1}{nh_{opt}\sqrt{2\pi}} \sum_{i=1}^{n} \exp\left(-\frac{(x - x_i)^2}{2h_{opt}^2}\right).$$

To assess the accuracy of the KDE estimate, $\hat{p}_{\mathrm{KDE}}(x)$, we compare it with the true underlying probability density function, $p(x) \sim \mathcal{N}(\mu, \sigma^2)$. The Mean Absolute Error (MAE) is then defined as

$$\mathrm{MAE} = \frac{1}{M} \sum_{j=1}^{M} \left|\hat{p}_{\mathrm{KDE}}(x_j) - p(x_j)\right|,$$

where $\{x_j\}_{j=1}^{M}$ are evaluation points.

A MAE of zero indicates that the two distributions are identical at the compared points. Typically a MAE in the range $[0, 1]$ is considered small, although this is dependent on the scale. For example, if two Gaussian distributions peak at 1, an MAE of 0.1 means that, on average, the absolute difference between their values at corresponding points is 0.1, or $\frac{0.1}{1} \times 100 = 10\%$.

Calculating this for a one dimensional standard Gaussian case (parameters: $\mu = 0.0$, $\sigma = 1.0$ and $n = 1000$). The estimated KDE and CV graphs are presented in Figure 2. The MAE is calculated to be 0.015503 corresponding to $\frac{0.015503}{0.4} \times 100 \approx 3.9\%$, indicating good agreement between the KDE and empirical PDFs.

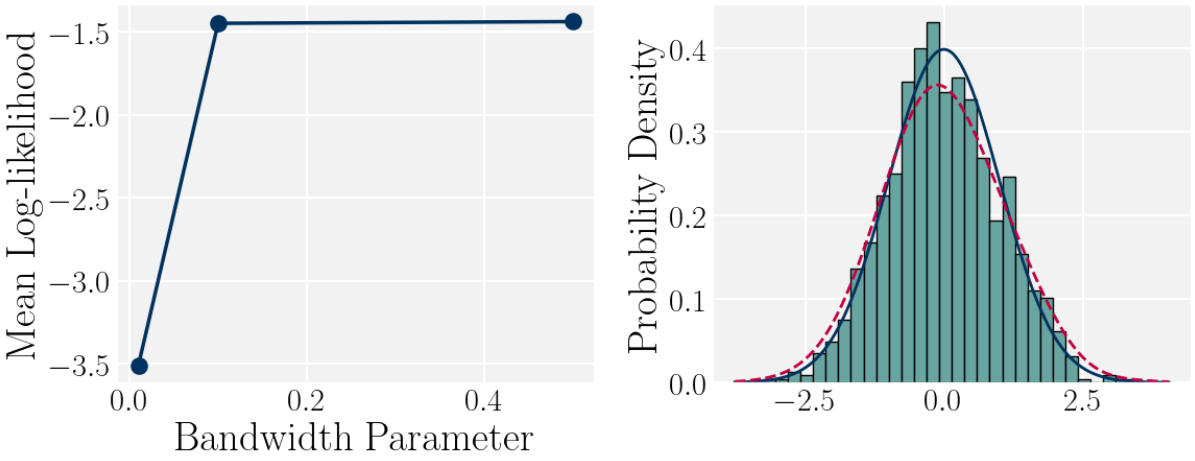

Figure 2: Cross-Validation and KDE Estimation Results. Left: CV score (mean log-likelihood) against bandwidth parameter. Right: KDE estimate (dashed pink) and empirical PDF (solid blue) plotted on top of the data histogram. The MAE is calculated as 0.015503.

In the multivariate setting, smoothing is controlled by a bandwidth matrix, $\mathbf{H}$. There is no unique method for determining $\mathbf{H}$, but a common simplifying assumption is that the dimensions are independent. This amounts to neglecting correlations between variables and assigning an individual bandwidth to each dimension, so that all off-diagonal elements vanish. Under this assumption, the problem reduces to a straightforward extension of the univariate case. An initial estimate for the i-th diagonal element of $\mathbf{H}$ can then be obtained using Silverman's Rule of Thumb [15]:

$$h_{\text{initial},i} = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} n^{-\frac{1}{d+4}} \sigma_i,$$

where σ_i is the standard deviation of the i-th dimension, d is the total number of dimensions, and n is the sample size. This provides an initial diagonal matrix $\mathbf{H}$ with entries $\mathbf{H_{ii}} = h_i^2$. To refine this estimate, cross-validation is performed over a range of scaling factors (typically within $[0.01, 1.5]$) applied to each h_i, optimising each dimension independently to yield the optimal values $h_{\text{opt},i}$. The final diagonal bandwidth matrix is

$$\mathbf{H} = \begin{bmatrix} h_{\text{opt},1}^2 & 0 & \cdots & 0 \\ 0 & h_{\text{opt},2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_{\text{opt},d}^2 \end{bmatrix}.$$

Applying this method to a standard two dimensional Gaussian with $n = 10000$ yields the results shown in Figure 3, achieving a MAE of 0.000617, indicating excellent agreement between the KDE and analytical PDF in two dimensions.

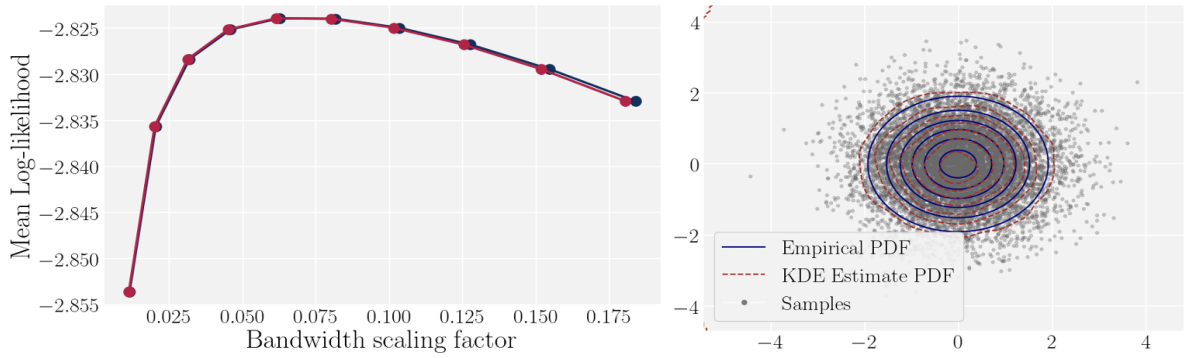

Figure 3: CV and KDE results in two dimensions. Left: mean log-likelihood (CV score) for the bandwidth scaling factors. Right: KDE estimated PDF (dashed pink), empirical PDF (solid blue), and sample data points (grey). The data has mean $\mu = [0, 0]$ and covariance matrix, $\mathbf{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

The marginal distributions which create this two dimensional Gaussian are shown in Figure 4.

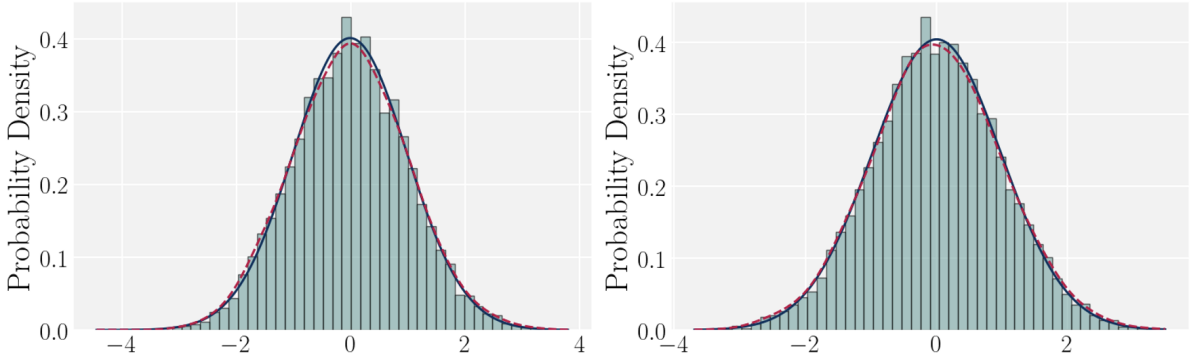

Figure 4: One dimensional marginal distributions of the previous two dimensional Gaussian distribution. Both distributions follow $\sim \mathcal{N}(0, 1)$ and are plotted with the KDE estimated PDF (dashed pink) and empirical PDF (solid blue). Both have MAE = 0.002656.

While assuming independence between dimensions may be effective in some cases, it is typically an unrealistic assumption for most multivariate distributions. Therefore we must use an alternative method to generate the bandwidth matrix that includes off-diagonal elements to capture the covariance. One method to achieve this is the Smooth Cross-Validation (SCV) technique [16].

SCV works by making an initial guess for the bandwidth matrix, with elements given by

$$\sqrt{H_{ij}} = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} n^{-\frac{1}{d+4}} \sigma_{ij},$$

where all symbols have their pre-defined meanings. Importantly, we no longer assume the off-diagonal elements to be zero.

From this initial matrix, we perform a Cholesky decomposition [17] to express the bandwidth matrix as $\mathbf{H} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is a lower-triangular matrix. This decomposition guarantees that $\mathbf{H}$ remains positive definite during optimisation. The elements of $\mathbf{L}$ are then optimised to minimise the SCV criterion; this criterion is set dependent on the problem. In our case, the SCV criterion is the negative mean log-likelihood of the KDE, because maximising the mean log-likelihood is equivalent to minimising the negative mean log-likelihood. Once $\mathbf{L}$ is optimised the full bandwidth matrix is reconstructed by $\mathbf{H} = \mathbf{L}\mathbf{L}^T$ which is symmetric and positive definite as required by KDE.

Completing this exercise with Gaussian data generated from mean $\mu = [0, 2]$ and covariance matrix, $\mathbf{\Sigma} = \begin{bmatrix} 1.7 & 2.3 \\ 0.5 & 0.2 \end{bmatrix}$ and $n = 10000$ gives Figure 5 with MAE $= 0.003314$

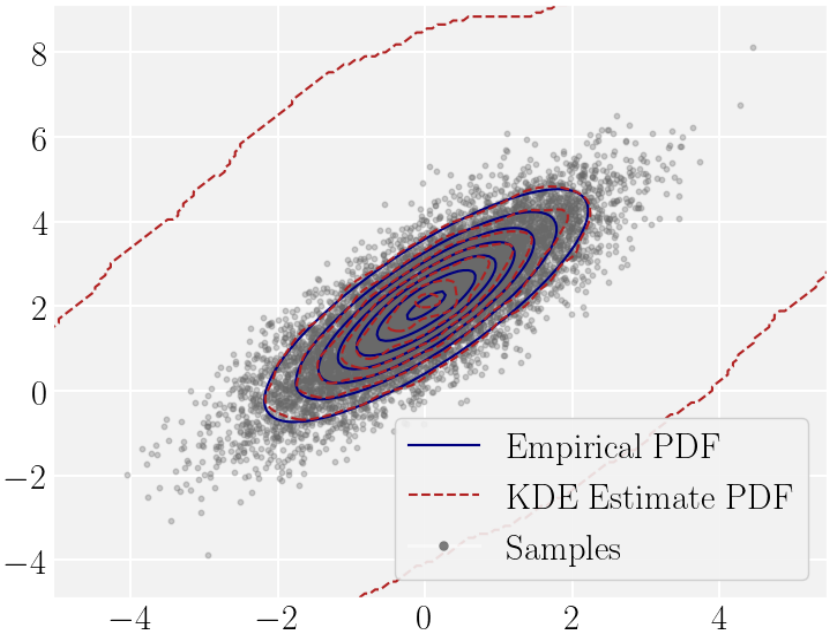

Figure 5: KDE estimated PDF (dashed pink), empirical PDF (solid blue), and sample data points (grey). The data has mean $\mu = [0, 2]$ and covariance matrix, $\mathbf{\Sigma} = \begin{bmatrix} 1.7 & 2.3 \\ 0.5 & 0.2 \end{bmatrix}$. MAE between the KDE and empirical PDF is 0.003314.

Throughout this analysis we have compared the KDE estimate with the true PDF, determined by known parameters for μ and $\mathbf{\Sigma}$. Clearly when applying this to data, we do not know these parameters, therefore the true PDF is replaced by the empirical PDF where

$$\hat{p}_{\text{emp}}(\mathbf{x}) \sim \mathcal{N}(\bar{\mathbf{x}}, \mathbf{S}),$$

where $\bar{\mathbf{x}}$ is the sample mean and $\mathbf{S}$ is the sample covariance matrix calculated from the data.

In summary, we have obtained two estimated probability density functions: the KDE estimated PDF and the empirical PDF. These can be used to compute statistical moments independently, and the difference between them indicates the accuracy of the KDE estimated PDF. However, before computing moments, it is helpful to adopt a more informative measure than the MAE

to assess the similarity between the two distributions. Hence, we use the Kullback–Leibler divergence, which provides a more reliable comparison of the two PDFs.

3.3 Kullback Leibler (KL) Divergence for Distribution Comparison

Up to this point we have been using MAE error to quantify how well the distributions fit together. This is a simple, symmetric distance measure based on the average of absolute differences between two distributions. However, it treats all deviations equally and does not account for the structure of the distributions. For this reason we turn to another metric to assess how well two distributions fit each other, the Kullback Leibler (KL) divergence [18], given by

$$D_{\mathrm{KL}}(\hat{p}_{\mathrm{emp}} \parallel \hat{p}_{\mathrm{KDE}}) = \int_{\mathbb{R}^d} \hat{p}_{\mathrm{emp}}(\mathbf{x}) \log \frac{\hat{p}_{\mathrm{emp}}(\mathbf{x})}{\hat{p}_{\mathrm{KDE}}(\mathbf{x})} \, d\mathbf{x}.$$

It is not a true distance metric because it is asymmetric, i.e. $D_{\mathrm{KL}}(\hat{p}_{\mathrm{emp}} \parallel \hat{p}_{\mathrm{KDE}}) \neq D_{\mathrm{KL}}(\hat{p}_{\mathrm{KDE}} \parallel \hat{p}_{\mathrm{emp}})$ and it does not satisfy the triangle inequality. Instead the KL divergence quantifies the difference between two probability distributions by measuring the expected logarithmic difference between their probabilities, weighted by the true distribution $\hat{p}_{\mathrm{emp}}$. For instance, if the distributions $\hat{p}_{\mathrm{emp}}(\mathbf{x})$ and $\hat{p}_{\mathrm{KDE}}(\mathbf{x})$ are close at a point $\mathbf{x}$, the ratio $\frac{\hat{p}_{\mathrm{emp}}(\mathbf{x})}{\hat{p}_{\mathrm{KDE}}(\mathbf{x})} \approx 1$ and thus $\log \frac{\hat{p}_{\mathrm{emp}}(\mathbf{x})}{\hat{p}_{\mathrm{KDE}}(\mathbf{x})} \approx 0$, which contributes little to the divergence. Likewise, if $\hat{p}_{\mathrm{emp}}(\mathbf{x})$ and $\hat{p}_{\mathrm{KDE}}(\mathbf{x})$ differ significantly, the ratio deviates from one and the logarithm term increases in magnitude, leading to a larger contribution to the divergence. When $\hat{p}_{\mathrm{KDE}}(\mathbf{x})$ underestimates $\hat{p}_{\mathrm{emp}}(\mathbf{x})$ (i.e. $\hat{p}_{\mathrm{KDE}}(\mathbf{x}) \ll \hat{p}_{\mathrm{emp}}(x)$), the logarithm becomes large and positive, causing a strong penalty. Overestimation ($\hat{p}_{\mathrm{KDE}}(\mathbf{x}) > \hat{p}_{\mathrm{emp}}(\mathbf{x})$) results in smaller penalties, reflecting the fact KL divergence is asymmetric. A divergence of zero would mean the two distributions are the same.

We now apply this metric to the two-dimensional Gaussian distribution from before (mean $\mu = [0, 2]$ and covariance matrix, $\mathbf{\Sigma} = \begin{bmatrix} 1.7 & 2.3 \\ 0.5 & 0.2 \end{bmatrix}$ and sample size, $n = 10000$). The two-dimensional KL divergence between the KDE estimate and the empirical PDF is computed using equation above. The integral is approximated via a Riemann sum over a uniform grid spanning from the minimum to the maximum values of the data along each axis. This procedure gives $D_{\mathrm{KL}} = 0.010064$.

Having quantified the closeness of the KDE to the true distribution using KL divergence, we now turn to computing statistical moments from the KDE estimated density. These provide insight into the structure and shape of the distribution.

3.4 Statistical Moments

Statistical moments are quantitative measures that describe the shape of a distribution. They can be defined in various ways with subtle differences. The simplest moments are raw moments which are defined through the moment generating function, $M_X(t)$ as

$$M_X(t) = \mathbb{E}[e^{tX}].$$

where the k-th raw moment is obtained by differentiating the moment generating function k times and evaluating at $t = 0$ like

$$\mu'_k = \left. \frac{d^k M_X(t)}{dt^k} \right|_{t=0}.$$

The zeroth raw moment is defined as

$$\mu'_0 = \mathbb{E}[X^0] = \mathbb{E}[1] = 1.$$

This simply reflects the fact that the total probability under a properly normalised PDF is 1. While it carries no shape information, in practise it is useful for verifying that a distribution is correctly normalised.

A major drawback of raw moments is they are calculated about the origin, without considering the distribution's location. To address this, central moments are defined about the mean, providing a more direct measure of the distribution's spread and shape. However, both raw and central moments depend on the scale of the distribution. To allow for comparison between distributions of different scales, central moments can be normalised by dividing by the standard deviation raised to the relevant power. These are called standardised moments, and are the most useful for us.

For our purposes, we are most interested the second centralised moment, the variance which tells us how a distribution varies in shape from the mean and the fourth standardised moment, known as the excess kurtosis. This measures the heaviness of the tails relative to a Gaussian distribution. A value of zero indicates Gaussian-like tails, thus excess kurtosis can be interpreted as how Gaussian a distribution is. A Gaussian-like distribution will have an excess kurtosis within the interval $[-3, 3$. It is calculated by

$$\text{Excess Kurtosis} = \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4} - 3.$$

The moments discussed above are computed for individual probability distributions. However, it is often useful to quantify how two distributions vary together. Let X and Y be two one-dimensional probability distributions. Their covariance is defined as

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])],$$

which measures the joint variability between the two distributions. To standardise this measure and make it easier to compare, we use the correlation, defined as

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \, \sigma_Y},$$

where $\sigma_X = \sqrt{\text{Var}(X)}$ and $\sigma_Y = \sqrt{\text{Var}(Y)}$ are the standard deviations of X and Y, respectively. We now turn to calculating these moments from our KDE-estimated PDFs.

3.5 Combining Statistical Moments and KDE

Having defined moments, we now link these concepts to KDE. From KDE we obtained an estimator $\hat{p}_{KDE}(\mathbf{x})$ from which these moments can be computed as integrals over the estimated density like

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

where $\{\mathbf{x}_i\}_{i=1}^n$ are the data points used in the KDE and $f(\mathbf{x}) = \mathbf{x}^k$ is the function representing the k-th moment. For example, the first moment (the mean), setting $f(\mathbf{x}) = \mathbf{x}$, we have

$$\mathbb{E}_{\hat{p}}[\mathbf{x}] = \int \mathbf{x}\hat{p}_{KDE}(\mathbf{x})\, d\mathbf{x} \approx \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i.$$

The equation above gives us two methods of calculating the moments that in theory give the same results, enabling us to compare and validate the KDE estimate. The right-hand side corresponds to calculating moments directly by averaging appropriate powers of the data, which we refer to as the empirical moment. The left-hand side uses numerical integration of the KDE estimate, which can be performed with methods such as grid-based integration or Monte Carlo techniques. For computational efficiency, we select Monte Carlo importance sampling.

3.6 Monte Carlo Importance Sampling

Monte Carlo methods are a broad class of algorithms that rely on random sampling to obtain numerical results [19]. Here we focus on one method, importance sampling.

Importance sampling aims to compute the expectation over any complicated target probability density function, $p(\theta)$ by actually calculating the expectation over a proxy probability density function, $p'(\theta)$ that is selected to match key features of the actual distribution [15]. Then to correct for the mismatch between these two distributions, each sample is multiplied by the importance weight, $\frac{p(\theta)}{p'(\theta)}$.

In our case, the target density is the KDE estimate $\hat{p}_{KDE}(x)$, and the proxy is a Gaussian distribution parametrised by the sample mean and covariance, $\hat{p}_{\text{emp}}(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}_{\text{sample}}, \boldsymbol{\Sigma}_{\text{sample}})$. Under this notation, the formula for importance sampling can be written as

$$\int f(\mathbf{x})\hat{p}_{\text{KDE}}(\mathbf{x})\, d\mathbf{x} = \int \left[f(\mathbf{x})\frac{\hat{p}_{\text{KDE}}(\mathbf{x})}{\hat{p}_{emp}(\mathbf{x})} \right] \hat{p}_{emp}(\mathbf{x})\, d\mathbf{x} \approx \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i)\frac{\hat{p}_{\text{KDE}}(\mathbf{x}_i)}{\hat{p}_{emp}(\mathbf{x_i})}.$$

where $f(\mathbf{x}) = \mathbf{x}^k$ is the function representing the k-th moment. (e.g., $f(\mathbf{x}) = \mathbf{x}$ for the mean).

As the name suggests, the ratio $\frac{\hat{p}_{KDE}}{\hat{p}_{emp}(x)}$ quantifies the relative contribution, or importance, of a sample x drawn from the proxy distribution $\hat{p}_{emp}(x)$ with respect to the target distribution $\hat{p}_{KDE}(x_i)$. By the Law of Large Numbers, importance sampling yields a consistent estimator. That is, as the number of samples $n \to \infty$, the weighted average

$$\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i)\frac{\hat{p}_{\text{KDE}}(\mathbf{x}_i)}{\hat{p}_{\text{emp}}(\mathbf{x}_i)},$$

converges in probability to the true expectation

$$\int f(\mathbf{x})\hat{p}_{\text{KDE}}(\mathbf{x})\, d\mathbf{x}.$$

This concludes the mathematical background underlying the project. We now move on to outline our methodology and assumptions.

4 Methodology

As mentioned, we use data from the NNPDF4.0 [11]. Our selected dataset consists of two bases: flavour and evolution. Each basis contains 1000 replicas (each replica is a parton distribution function) evaluated at 50 grid points for each of nine chosen parton flavours (up, down, strange, charm, anti-up, anti-down, anti-strange, anti-charm quarks and gluon). This results in a 450-dimensional grid of replicas.

As we wish to use KDE estimation to reconstruct the parton distribution functions, we begin by checking how Gaussian the underlying one dimensional replicas are by calculating the excess kurtosis for each of the 450 one dimensional replicas. As we find the majority of replicas display Gaussian behaviour, we proceed to calculate the one and two dimensional KDE distributions then use these to reconstruct the 450 dimensional covariance matrix. This is achieved by looping through all possible two dimensional combinations of flavour basis, then calculate two dimensional matrices and construct the full matrix by averaging where elements overlap. This is then normalised to give the correlation matrix. Where appropriate we provide quantitative errors which are all calculated using bootstrap sampling.

We are using Gaussian distributions in two places. First, in KDE we use a Gaussian kernel. Secondly in importance sampling we employ a Gaussian proxy distribution parametrised by the sample mean and sample covariance. Although the Gaussian assumption is not perfectly satisfied in all cases, we demonstrate that it holds sufficiently for these methods to remain valid.

When constructing 2D bandwidth matrices we enforce positive semi-definiteness. Samples producing non-positive semi-definite bandwidth matrices are excluded. Additionally, to maintain numerical stability, all bandwidth matrix elements are bounded below by 10^{-9}. This threshold addresses numerical instability observed at higher grid indices (approximately index 45 and above), where parton distribution function constraints produce extremely small values.

With these data characteristics and assumptions established, we proceed to present results derived from this method.

5 Results and Discussion

5.1 Analysis of Excess Kurtosis

To investigate the Gaussianity of the whole dataset, we loop through all 450 dimensions and calculate the excess kurtosis for each replica. This is visualised as a histogram in Figure 6. Theoretically parton distribution function are constrained to vanish when $x = 1$, corresponding to the 50-th grid point. Therefore the parton distribution function cannot fluctuate in any arbitrary manner near the 50-th grid point. In practise this leads to small values for high grid indices ($\sim 10^{-45}$) which gives large excess kurtosis values (~ 150). These are excluded from Figure 6 by enforcing an upper limit of five.

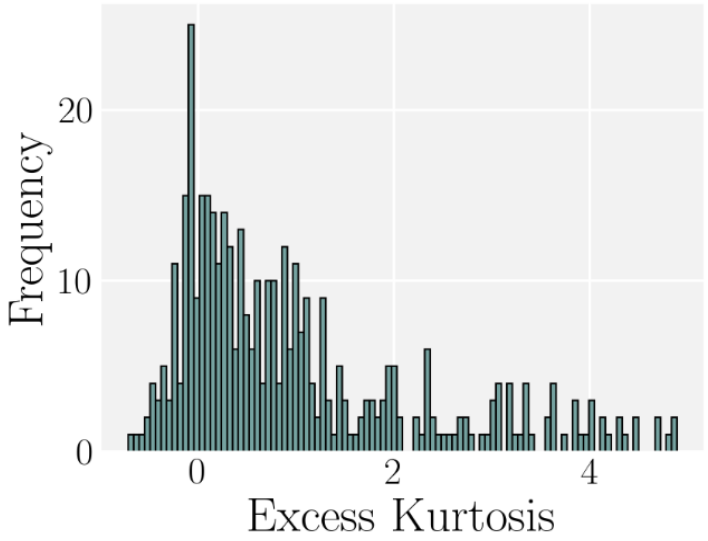

Figure 6: Excess Kurtosis Histogram calculated through empirical calculation. An upper limit of five is imposed.

Recall excess kurtosis acts as a measure of how Gaussian a distribution is and we categorise a large excess kurtosis as greater than three. We can see the majority of replicas fit within the range $[-3, 3]$ suggesting the majority of replicas are consistent with Gaussian behaviour. although we note a significant number that exhibit deviations from Gaussian behaviour.

In order to diagnose which flavours or grid indices demonstrate the worst agreement, we plot the excess kurtosis against grid index for each flavour. Figure 7 shows this for all flavours.

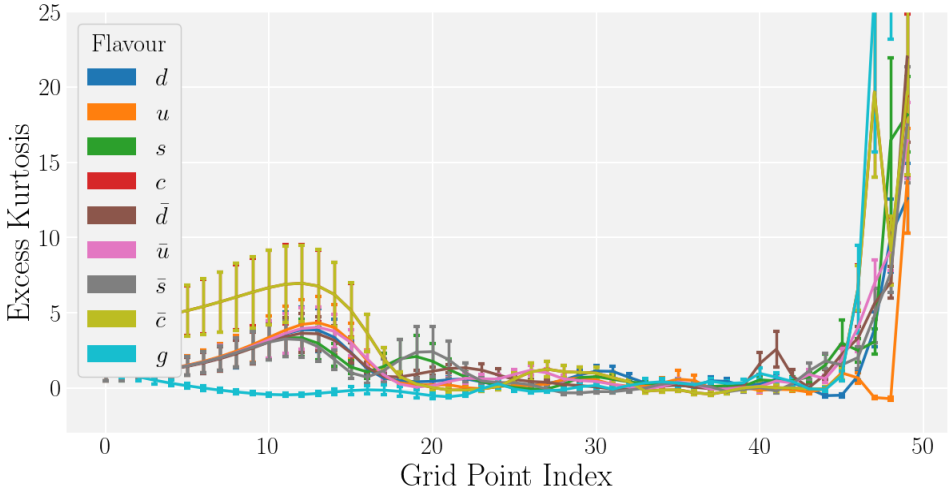

Figure 7: Excess Kurtosis (calculated empirically) plotted against grid index for all nine parton flavours.

From Figure 7 we can see the all flavours follow a similar trend, with replicas being approximately Gaussian within error upto index ~ 45, after this point the deviation from Gaussianity is clear. Due to these deviations, we limit our analysis to 45 grid points, as the assumption of Gaussianity clearly does not hold beyond this point. Despite some departures from Gaussianity,

the majority of replicas show significant agreement, satisfying our assumptions and allowing us to proceed with KDE reconstructions.

5.2 One Dimensional KDE Reconstruction

We randomly select the 28th grid point to examine the distributions of the up quark and gluon. The resulting probability density functions are shown in Figure 8.

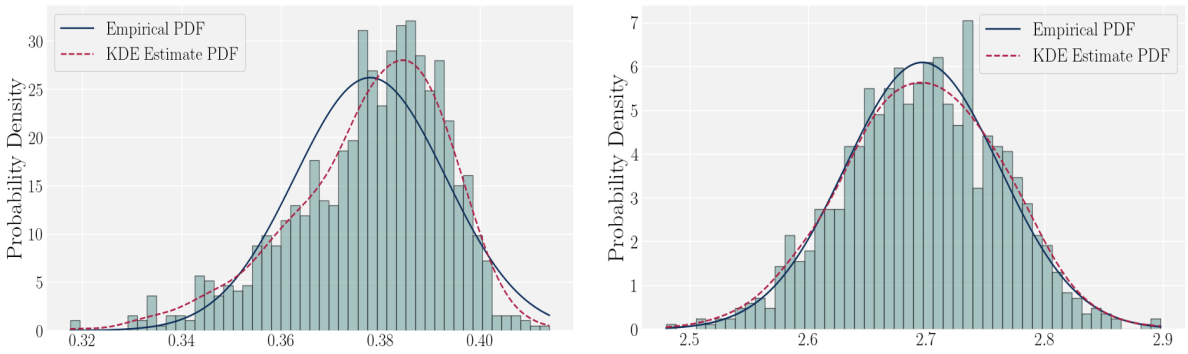

Figure 8: KDE PDF (dashed pink) at grid point 28 with empirical PDF (solid blue). Left: up quark with $D_{KL} = 0.036490$. Right: gluon with $D_{KL} = 0.003270$.

With KL divergence values of 0.036490 and 0.003270 for the up quark and gluon respectively. The higher KL divergence for the up quark reflects the fact it is clearly non-Gaussian, whereas the gluon shows a distribution that resembles a Gaussian. We now proceed to calculate the zeroth moment, mean and variance to see how these are effected by this non-Gaussian behaviour. The results are summarised in Table 1

Quantity	**Up quark**	**Gluon**
	KDE	
Zeroth Moment	$1.0000 \pm 4.8682 \times 10^{-4}$	$1.0000 \pm 2.2063 \times 10^{-3}$
Expectation	$0.37791 \pm 4.8682 \times 10^{-4}$	$2.6965 \pm 2.20628 \times 10^{-3}$
Variance	$2.5483 \times 10^{-4} \pm 1.1647 \times 10^{-5}$	$4.7098 \times 10^{-3} \pm 1.9261 \times 10^{-4}$
	Empirical	
Expectation	0.37791 ± 1.1650	$2.6965 \pm 2.2063 \times 10^{-3}$
Variance	$2.3187 \times 10^{-4} \pm 2.0280 \times 10^{-3}$	$4.2855 \times 10^{-3} \pm 2.0506 \times 10^{-3}$

Table 1: Summary of statistical moments for two one dimensional distributions: up quark and gluon, with errors included (rounded to five significant figures).

From the calculations we can see the up quark KDE estimation deviates much more than the gluon, this is likely a reflection of the non-Gaussian nature of the up quark distribution. Altogether, the error in variance values is greater than in the expectation values. This reflects the numerical error introduced by squaring the expectation values, which amplifies values that are already small to begin with, thereby increasing numerical instability. Overall these observations demonstrate cases where the one dimensional replicas can be non-Gaussian, introducing more error into our moment calculations. We now proceed to the two dimensional KDE case.

5.3 Two Dimensional KDE Reconstruction

Having justified that a Gaussian approximation is reasonable upto the 45th grid index, we now extend the analysis to higher-dimensional KDE estimates. Using the same down quark and

gluon at the 28th grid point, we compute the two-dimensional KDE, shown in Figure 9.

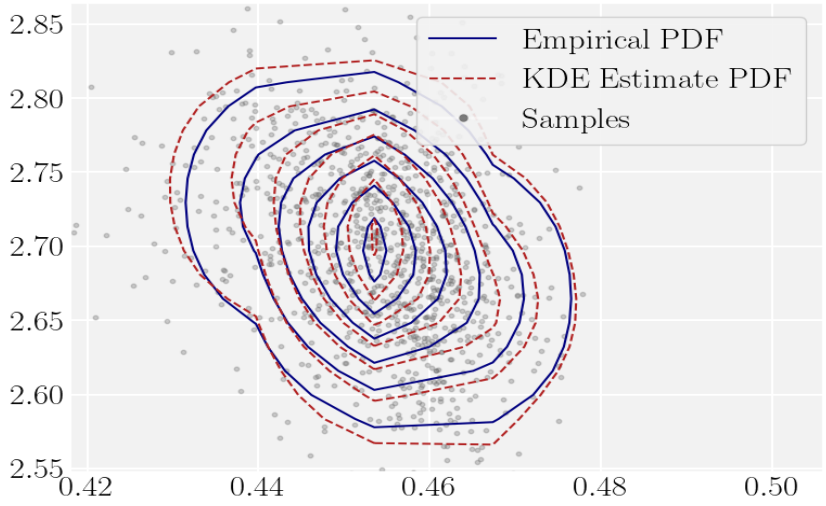

Figure 9: Two-dimensional KDE estimate for up quark and gluon at grid point 28. KL divergence: $D_{KL} = 0.040222$.

The KL divergence is calculated to be 0.040222 which indicates the KDE estimate is in good agreement with the empirical PDF. The asymmetry observed on one side reflects the skewness already noted in the one-dimensional up quark distribution. With this KDE estimate we calculate the zeroth moment, expectation, variance and covariance with associated errors, the results are presented in Table 2.

Quantity	**Up quark**	**Gluon**
KDE		
Zeroth Moment	0.93258 ± 0.22898	
Expectation	0.45249 ± 0.002895	2.6868 ± 0.019538
Variance	$6.9729 \times 10^{-5} \pm 3.1320 \times 10^{-5}$	$2.7401 \times 10^{-3} \pm 1.4458$
Covariance	$-1.5177 \times 10^{-4} \pm 1.3423 \times 10^{-4}$	
Empirical		
Expectation	0.45408 ± 0.00029216	2.6965 ± 0.002053
Variance	$1.0041 \times 10^{-4} \pm 5.2367 \times 10^{-6}$	$4.2855 \times 10^{-3} \pm 1.8510 \times 10^{-4}$
Covariance	$-2.1531 \times 10^{-4} \pm 5.2367 \times 10^{-6}$	

Table 2: Summary of statistical moments for two dimensional distribution made of two flavours (up quark and gluon), with errors included (rounded to five significant figures).

In these results, the KDE moments agree with the empirical moments within error although the agreement is noticeable worse than the one dimensional case. We now proceed to construct the full covariance and correlation matrices.

5.4 Global Covariance Matrix Reconstruction

Directly computing the 450-dimensional covariance matrix is computationally infeasible, and the accuracy of KDE deteriorates significantly for dimensions higher than two. To address this, we reconstruct the covariance matrix using two-dimensional KDE estimates for each pair of parton flavours. Furthermore, our excess kurtosis analysis indicates that Gaussianity does not hold

beyond grid index 45; therefore, we exclude the final five indices for each flavour. The matrix is constructed by iterating over all combinations of the 45 grid points and 9 flavours, estimating the two-dimensional KDE for each pair. The global covariance matrix is then obtained by averaging overlapping entries across these estimates.

Figure 10 shows the reconstructed correlation matrices for all nine flavours and the first 45 grid points giving overall dimensions 405×405.

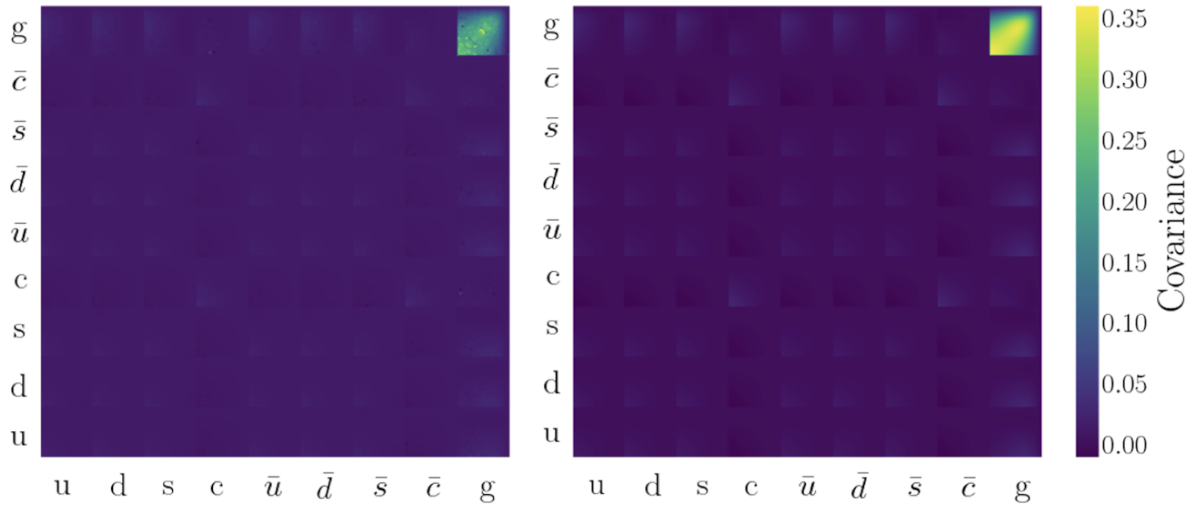

Figure 10: Comparison of 45 grid point correlation matrices constructed using 2D KDE reconstruction (left) and empirical data (right). Each visible block corresponds to correlations between grid points for a specific pair of parton flavours, ordered as $u, d, s, c, \bar{d}, \bar{u}, \bar{s}, \bar{c}, g$ (left to right, bottom to top). Both methods show consistent overall structure, but the scale is dominated by gluon-gluon variance. Dimensions 405×405.

We see both matrices demonstrate high covariance for the gluon pair which dominates the scale, interestingly, while the empirical covariance matrix exhibits a strong, high-covariance diagonal, the KDE-based covariance matrix shows lower covariance values and a much less pronounced diagonal peak for this pair. It is difficult to tell any interesting features for the quarks due to the high variance of the gluon which dominates the scale. In order to combat this, we calculate the correlation, which is effectively the covariance normalised.

5.5 Global Correlation Matrix Reconstruction

The correlation is calculated similarly to the covariance, differing only by a normalisation factor applied at the end. We perform this reconstruction for the same three sizes. Figure 11 shows the reconstructed correlation matrices for all nine flavours and the first 45 grid points, corresponding to total dimensions of 405×405.

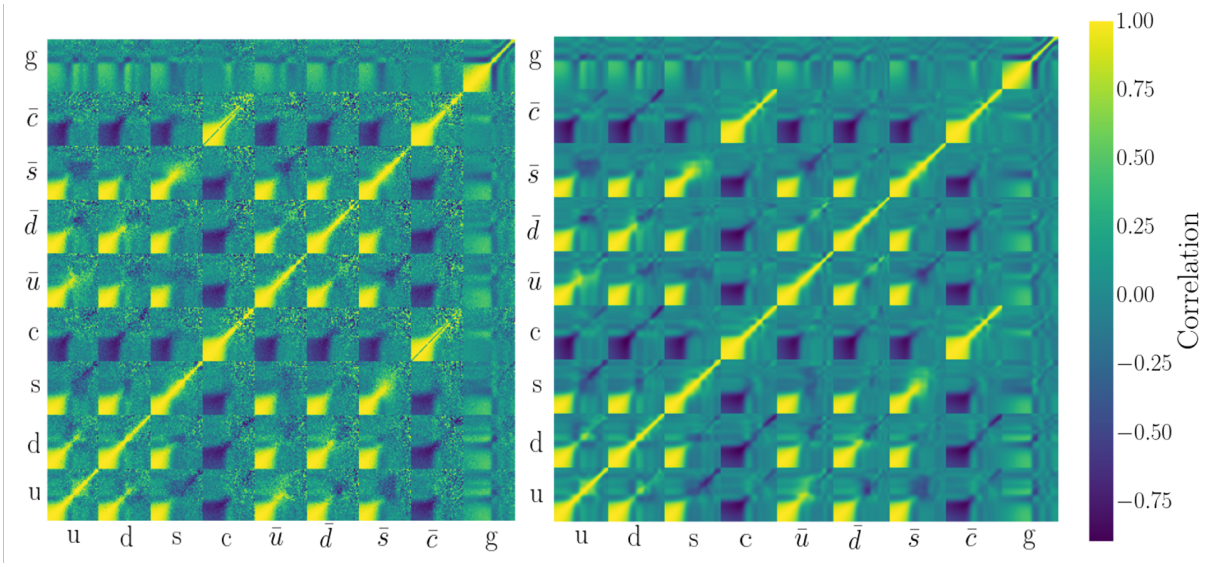

Figure 11: Comparison of 45 grid point correlation matrices constructed using empirical data (right) and 2D KDE reconstruction (left). Each visible block corresponds to correlations between grid points for a specific pair of parton flavours, ordered as $u, d, s, c, \bar{d}, \bar{u}, \bar{s}, \bar{c}, g$ (left to right, bottom to top). Both methods show consistent overall structure, but the KDE reconstruction appears more grainy and less sharply defined. Dimensions 405×405.

By inspection, the diagonal elements corresponding to the same parton (e.g., gluon–gluon) are all equal to 1, as expected. This provides a useful self-correlation check, confirming that our method is functioning correctly. In addition, for each parton pair (represented by a smaller square), we observe the following patterns. All flavours except charm, anti-charm, and gluon show a positive correlation when both partons in the pair have low grid indices. If one parton has a low index and the other a high index (or vice versa), the correlation is near zero. When both partons have high indices, the correlation can vary; sometimes higher and sometimes lower depending on the specific pair. Charm and anti-charm partons follow a similar pattern, except when both have low grid indices, where there is a strong anti-correlation (with the exception of the charm, anti-charm pair). Gluons generally exhibit low correlations with all other partons, roughly ranging from -0.25 to 0.25, except for a strong correlation with themselves (top right corner square). While sum laws and physical constraints like momentum sum rule, valence sum rules, small-x behaviour, and positivity [20] induce correlations it is not clear that these are responsible for the pattern we see here. Further work is needed to determine the physical origin of this pattern.

When comparing the KDE and the empirical matrices, we observe that the main pattern is generally consistent between the two. However, the KDE appears noticeably grainy in regions outside the main pattern within each square, particularly for higher grid indices. These deviations are likely due to our assumptions (such as Gaussianity) and numerical approximation errors. Overall, these results demonstrate that the KDE provides a reliable reconstruction of the empirical patterns

6 Conclusions and Future Work

In conclusion, we have demonstrated that constructing global correlation matrices from two-dimensional KDE approximations of correlations is effective. This is first evidenced by assessing the Gaussianity of the NNPDF4.0 replica ensembles. By calculating the excess kurtosis, we find that most marginal distributions are approximately Gaussian. Specifically, the excess

kurtosis for the marginals remains within the Gaussian range ($[-3, 3]$) for all grid indices up to approximately index 45, reflecting the constraint that the parton distribution function vanishes at the 50-th grid point.

Next, we reconstruct the PDF in one dimension. The up quark at the 28th grid index exhibits some deviations from Gaussian behaviour, whereas the gluon at the same index remains Gaussian. One-dimensional statistical moments are calculated and found to agree with empirical estimates within errors. The deviations observed in the up quark are attributed to its reduced Gaussianity. This analysis is then extended to two dimensions using the same replica ensembles, and the statistical moments again show agreement with empirical estimates within error.

Finally, we employ two-dimensional KDE to calculate statistical moments and find good agreement with empirical estimates. Two-dimensional covariance estimates are then combined and normalized to construct a global correlation matrix. The resulting matrix exhibits distinct patterns consistent with the empirical matrix, though additional work is required to determine the errors between the empirical and KDE-derived correlations and to understand the physical origin of these patterns.

Future work could focus on calculating observable quantities, such as cross sections or structure functions, and comparing them with experimental results to investigate how Gaussian assumptions propagate into these quantities. Additionally, improvements in computational efficiency could be explored, for instance through high-performance computing techniques, to reduce the computation time required for these higher-dimensional matrices.

References

1. Butterworth, J. & et al. PDF4LHC recommendations for LHC Run II. *J. Phys. G* (2016).
2. Rojo, J. *et al.* The PDF4LHC report on PDFs and LHC data: results from Run I and preparation for Run II. *Journal of Physics G: Nuclear and Particle Physics* **42,** 103103. ISSN: 1361-6471. http://dx.doi.org/10.1088/0954-3899/42/10/103103 (Sept. 2015).
3. Ball, R. D. & et al. A determination of parton distributions with faithful uncertainty estimation. *Nucl. Phys. B* (2009).
4. Ball, R. D. & et al. Parton distributions for the LHC Run II. *JHEP* (2015).
5. Ball, R. D. & et al. Precision determination of electroweak parameters and parton distributions with NNPDF3.0. *Eur. Phys. J. C* (2017).
6. Ball, R. D. & et al. The path to proton structure at 1% accuracy. *Eur. Phys. J. C* (2022).
7. Dulat, S. & et al. New parton distribution functions from a global analysis of quantum chromodynamics. *Phys. Rev. D* (2016).
8. Hou, T. & et al. New CTEQ global analysis of quantum chromodynamics with high-precision data from the LHC. *Phys. Rev. D* (2021).
9. Bailey, S., Cridge, T., Harland-Lang, L. A., Martin, A. D. & Thorne, R. S. Parton distributions from LHC, HERA, Tevatron and fixed target data: MSHT20 PDFs. *The European Physical Journal C.* http://dx.doi.org/10.1140/epjc/s10052-021-09057-0 (2021).
10. Ball, R. D. & et al. Unbiased global determination of parton distributions and their uncertainties at NNLO and LO. *Nucl. Phys. B* (2012).
11. Iranipour, S. & Ubiali, M. A new generation of simultaneous fits to LHC data using deep learning. *JHEP* (2022).
12. Carrazza, S., Forte, S., Kassabov, Z., Latorre, J. I. & Rojo, J. An unbiased Hessian representation for Monte Carlo PDFs. *Eur. Phys. J. C* **75,** 369 (2015).
13. Coughlan, G. D. & Dodd, J. E. *The ideas of particle physics: An introduction for scientists* (Cambridge University Press, 2023).
14. Feynman, R. P. The behavior of hadron collisions at extreme energies. *Conference Proceedings C* (1969).
15. Sugiyama, M. *Introduction to statistical machine learning* (Morgan Kaufmann, 2016).
16. Golub, G. H. & Van Loan, C. F. *Matrix computations* (Johns Hopkins University Press, 2013).
17. Hall, P., Marron, J. S. & Park, B. U. Smoothed cross-validation. *Probability Theory and Related Fields* (1992).
18. MacKay, D. J. C. *Information theory, inference and learning algorithms* (Cambridge University Press, 2003).
19. Owen, A. B. *Monte Carlo theory, methods and examples* (https://artowen.su.domains/mc/, 2013).
20. Ball, R. D. *et al.* The Path to Proton Structure at One-Percent Accuracy. *arXiv preprint arXiv:2109.02653.* Accessed: 2025-08-22. https://arxiv.org/abs/2109.02653 (2021).