UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

Laurea Triennale in Fisica

Statistical properties of PDF uncertainties

Relatore:
Prof. Stefano Forte

Tesi di Laurea di:
Alessio Ghidini
Matricola: **11548A**

Anno Accademico 2024/2025

Contents

Introduction

Understanding the internal structure of hadrons represents one of the central themes in high-energy physics. Since the pioneering deep inelastic scattering (DIS) experiments conducted in the late 1960s, it has become evident that particles such as the proton and neutron possess an internal substructure composed of more fundamental constituents: quarks and gluons. These constituents, collectively known as partons, are described within the theoretical framework of Quantum Chromodynamics (QCD).

To characterize the distribution of partons within a hadron, one introduces the Parton Distribution Functions (PDF). These functions describe the probability of finding a parton of a given type carrying a specific fraction of the hadron's total momentum. PDF still cannot be computed directly from first principles in QCD; instead, they are determined empirically through the analysis and extrapolation of experimental data.

Due to the way the PDF are calculated, they inevitably carry uncertainties that must be quantified and propagated to ensure the reliability of physical predictions. To address this challenge, the NNPDF collaboration [1] has developed an approach based on machine learning techniques, in particular the use of neural networks (for a detailed description of their work, refer to [2]). This methodology allows for a flexible representation of the functional form of the PDFs and enables the direct propagation of experimental uncertainties within the model.

In this thesis, we investigate the Hessian conversion method, which allows the transformation of the uncertainties of a PDF set obtained through the NNPDF approach into the Hessian representation. In this formalism, uncertainties are assumed to follow a multivariate Gaussian distribution; consequently, the error is characterized by a set of eigenvector PDF pairs that define the boundaries of the confidence ellipsoid in the parameter space.

To quantify these uncertainties and carry out the analysis, the $\Delta\chi^2$ distribution is employed. This distribution measures the overall χ^2 variation with respect to its global minimum, χ^2_{min}, along the directions of the eigenvectors.

The main objective of this work is to assess the performance of the Hessian conversion method when applied to an NNPDF set.

The analysis presented in this thesis compares the results of the Hessian conversion applied to a standard PDF set with those obtained from a PDF generated through a closure test. The closure test serves as an internal validation procedure, in which a PDF is constructed from a known "underlying truth" distribution, thereby allowing the reliability of the method to be evaluated under controlled conditions.

The structure of this thesis is organized as follows:

- **Chapter 1** provides the theoretical foundation for understanding Parton Distribution Functions (PDFs). After introducing the main aspects of Quantum Chromodynamics (QCD), including the QCD Lagrangian, the running of the strong coupling, and the phenomena of asymptotic freedom and confinement, it discusses the interplay between perturbative and non-perturbative regimes. The need for PDFs emerges as an effective probabilistic description of parton momentum distributions inside hadrons. The chapter also discusses the formal definition of the PDFs, their relation to hadronic observables, and the factorization theorem that separates long-and short distance effects. It concludes by outlining higherorder QCD corrections and the evolution equations that govern PDF scale dependence.

- **Chapter 2** constitutes the methodological core of this thesis, presenting the theoretical and computational tools required for the subsequent analysis. It outlines the two main strategies used to extract Parton Distribution Functions (PDFs) from experimental data and quantify their associated uncertainties: the Hessian approach, based on functional parametrizations and quadratic error propagation, and the Monte Carlo approach, which relies on statistical sampling and neural network representations. The chapter also discusses the Hessian conversion procedure, which allows a consistent transformation between Monte Carlo and Hessian sets.

- **Chapter 3** presents the core results of this work, focusing on the quantitative analysis of the $\Delta\chi^2$ parameter and on the validation of the Hessian conversion methodology. The chapter begins by examining the global properties of the $\Delta\chi^2$ distribution, comparing its behavior in the case of genuine experimental data and in closure test environments. Subsequently, the dependence of the $\Delta\chi^2$ parameter on the number of Monte Carlo replicas is investigated through both quadratic and quartic polynomial fits. Finally, the chapter explores the dependence of the results on the number of eigenvalues used in the Hessian conversion.

Chapter 1

Theoretical Framework for Parton Distributions

Understanding how parton distribution functions (PDFs) describe the internal structure of the proton requires an introduction to the fundamental concepts of Quantum Field Theory, specifically Quantum Chromodynamics (QCD). For a more detailed treatment of this topic, readers are referred to specialized texts such as [3, 4].

1.1 Fundamentals of Quantum Chromodynamics (QCD)

1.1.1 The QCD Lagrangian

The proton's internal structure is constituted by quarks, whose interactions are governed by the strong force, as described by Quantum Chromodynamics. QCD is a non-abelian gauge theory with the gauge group SU(3) and is defined by its Lagrangian density, $\mathcal{L}_{\text{QCD}}$. A common effective form of this Lagrangian is:

$$\mathcal{L}_{\text{eff}}^{\text{QCD}} = \mathcal{L}_{\text{classic}} + \mathcal{L}_{\text{gauge}} + \mathcal{L}_{\text{ghost}}. \tag{1.1}$$

The first term, $\mathcal{L}_{\text{classic}}$, describes the interactions between quarks and gluons. The other two terms, $\mathcal{L}_{\text{gauge}}$ and $\mathcal{L}_{\text{ghost}}$, are necessary to make perturbative calculations possible. These terms are a result of gauge fixing, a procedure required because the classical Lagrangian's gauge invariance introduces an infinite number of field configurations for a single physical state, which complicates calculations.

The classical Lagrangian, in its original form as introduced by Yang and Mills, is given by:

$$\mathcal{L} = \sum_f \bar{\psi}_a (i\not{D} - m)_{ab} \psi_b - \frac{1}{4} F^a_{\mu\nu} F^{\mu\nu a}. \tag{1.2}$$

This Lagrangian describes the interactions of quark fields ψ^a (which transform under the fundamental representation of the SU(3) color group with $a = 1, 2, 3$) and gluon fields. The gluon fields are contained within the covariant derivative $\not{D}$ and the field strength tensor $F^a_{\mu\nu}$. Here, m is the quark mass parameter, and $\not{D}$ is the contraction of Dirac matrices with the covariant derivative, defined as $\not{D} = \gamma_\mu D^\mu = \gamma_\mu(\partial_\mu - igA_\mu)$, where γ_μ represent the Dirac matrices, A_μ is the gluon field and g is the bare coupling of the theory which determines the strength of the interaction. The field strength tensor $F^a_{\mu\nu}$ is defined in terms of the gluon vector field A^a_μ as:

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g f_{abc} A^b_\mu A^c_\nu. \tag{1.3}$$

In this expression, f_{abc} are the structure constants of the SU(3) Lie algebra.

1.1.2 Strong Coupling and Asymptotic Freedom

Analogous to what is done in Quantum Electrodynamics (QED), it is useful to define a strong coupling constant, α_s related to the g used in 1.3 by:

$$\alpha_s = \frac{g^2}{4\pi}. \tag{1.4}$$

The behavior of α_s provides crucial insights into the nature of the strong interaction. Due to QCD's non-abelian nature, the emission of a gluon changes the color charge of the emitting particle. This leads to a phenomenon called antiscreening, which is the opposite of the screening effect observed in QED. Mathematically, this is expressed by the dependence of α_s on the momentum transfer scale, Q^2 , in an interaction:

$$\frac{d\alpha_s(Q^2)}{dQ^2} < 0. \tag{1.5}$$

This indicates that the strong coupling $\alpha_s(Q^2)$ decreases as the momentum transfer Q^2 increases. This property, known as asymptotic freedom, allows for the use of perturbative methods in high-energy regimes.

Conversely, as the momentum transfer decreases (i.e., at low energies), the value of α_s grows significantly, rendering perturbative calculations unreliable. This behavior is consistent with confinement, the characteristic property of quarks that prevents them from being observed as free particles and is the environment probed by PDFs.

1.1.3 Renormalization Group and Scale Dependence

The dependence of α_s on the energy scale is described by the renormalization group. This formalism allows us to evaluate the coupling constant at different energy scales and, in turn, describe how PDFs evolve with these scales. The running of α_s can be derived by requiring that a physical quantity (e.g., a scattering amplitude) be independent of the arbitrary renormalization scale μ^2. At first order, this yields the well known solution for the running coupling:

$$\alpha_s(\mu^2) = \frac{\alpha_s(\mu_0^2)}{1 + \frac{\beta_0}{4\pi}\alpha_s(\mu_0^2)\ln\left(\frac{\mu^2}{\mu_0^2}\right)}, \tag{1.6}$$

where β_0 is the first coefficient of the beta function, a positive constant given by:

$$\beta_0 = \frac{11N_c - 2n_f}{3}. \tag{1.7}$$

Here, N_c is the number of colors ($N_c = 3$ for QCD) and n_f is the number of quark flavors. This equation demonstrates how the coupling constant changes with the energy scale, a fundamental result that underpins the theoretical framework for describing the evolution of parton distribution functions.

1.2 Deep Inelastic Scattering and Cross-Sections

The determination of parton distribution functions (PDFs) relies on experimental observables, specifically cross-sections. While various processes can measure cross-sections, we will focus on Deep Inelastic Scattering (DIS), a fundamental process for probing the internal structure of hadrons.

In a DIS experiment, an elementary lepton (l), such as an electron, muon, or neutrino, collides with an hadron (h), producing a generic final state X. This process is represented as:

$$l(k) + h(p) \to l'(k') + X, \tag{1.8}$$

The four-momentum transfer between the lepton and the hadron is $q^\mu = k^\mu - k'^\mu$, with the squared momentum transfer defined as $Q^2 = -q^2$.

The kinematics of the interaction are commonly described by the Bjorken scaling variable, x:

$$x = \frac{-q^2}{2p \cdot q} = \frac{Q^2}{2m_h\nu}. \tag{1.9}$$

Here, $\nu = E_l - E_{l'}$ is the energy transferred from the lepton to the hadron in the hadron's rest frame.

At the lowest order of the electroweak interaction, the differential cross-section

can be factorized into a leptonic part and a hadronic part:

$$d\sigma = \frac{d^3k'}{2s|\mathbf{k}'|}\frac{c_V^4}{4\pi^2(q^2-m_V^2)^2}L_{lV}^{\mu\nu}(k,q)W_{\mu\nu}^{Vh}(p,q). \tag{1.10}$$

In this expression, V denotes the exchanged vector boson (e.g., a photon γ or a weak boson $W^{\pm}, Z^0$) with mass m_V. The terms $L_{lV}^{\mu\nu}$ and $W_{\mu\nu}^{Vh}$ are the leptonic tensor and the hadronic tensor, respectively. The hadronic tensor is of particular interest as it contains all information about the hadron's structure. Due to symmetry properties, the hadronic tensor can be parameterized in terms of three structure functions, W_1, W_2, and W_3:

$$W_{\mu\nu}^{(Vh)} = \left(-g_{\mu\nu}+\frac{q_\mu q_\nu}{q^2}\right)W_1^{(Vh)}(x,Q^2) \tag{1.11}$$

$$+\left(p_\mu - q_\mu\frac{p\cdot q}{q^2}\right)\left(p_\nu - q_\nu\frac{p\cdot q}{q^2}\right)\frac{1}{m_h^2}W_2^{(Vh)}(x,Q^2) \tag{1.12}$$

$$-i\epsilon_{\mu\nu\lambda\sigma}p^\lambda q^\sigma\frac{1}{m_h^2}W_3^{(Vh)}(x,Q^2). \tag{1.13}$$

These functions are often replaced by the more commonly used structure functions F_i, defined as:

$$F_1(x,Q^2) = W_1(x,Q^2); \tag{1.14}$$

$$F_2(x,Q^2) = \frac{\nu}{m_h}W_2(x,Q^2); \tag{1.15}$$

$$F_3(x,Q^2) = \frac{\nu}{m_h}W_3(x,Q^2). \tag{1.16}$$

These structure functions are the direct observables measured in DIS experiments by analyzing the scattered lepton's momentum.

1.2.1 Connection between Cross-Sections and Parton Distributions

At leading order in perturbative QCD, the scattering of a nucleon is described as the incoherent sum of the elastic scattering of a lepton off the individual quarks and gluons that constitute the nucleon. This implies that the total cross-section is given by the probability of finding a parton with a certain momentum fraction, multiplied by the cross-section of the elementary lepton-parton scattering. This relationship can be expressed as:

$$\frac{d\sigma_{(lh)}}{dE_{k'}d\Omega_{k'}}(p,q) = \sum_f\int_0^1 d\xi\frac{d\sigma_{\text{Born}}^{(lf)}}{dE_{k'}d\Omega_{k'}}(\xi p,q)\phi_{f/h}(\xi), \tag{1.17}$$

where $\phi_{f/h}(\xi)$ is the parton distribution function (PDF), representing the probability of finding a parton of flavor f with momentum fraction ξ inside the hadron h. The term $\frac{d\sigma^{(lf)}_{\text{Born}}}{dE_{k'}d\Omega_{k'}}$ is the cross-section for the elastic scattering of a lepton off a quark, calculated from the Born diagram .

The structure functions can then be expressed in terms of a convolution of the partonic structure functions and the PDFs:

$$F_a^{(Vh)}(x) = \sum_f \int_0^1 \frac{d\xi}{\xi} F_a^{(Vf)}(x/\xi)\phi_{f/h}(\xi) \qquad (a = 1, 3); \tag{1.18}$$

$$F_2^{(Vh)}(x) = \sum_f \int_0^1 d\xi F_2^{(Vf)}(x/\xi)\phi_{f/h}(\xi). \tag{1.19}$$

The partonic structure functions, $F_i^{(Vf)}$, are calculable from the elementary interaction vertex.

For simplicity, let's summarize the results for charged-current DIS (with neutrino or antinuetrino beans), where the structure functions of neutrino (W^+) and antineutrino (W^-) interactions are combined into sums and differences:

$$F_{i\pm}^{(Wh)} = \frac{1}{2}(F_i^{W^+h} \pm F_i^{W^-h}). \tag{1.20}$$

By introducing the notation $U_h(x)$ for the distributions of charge $+2/3$ quarks (up, charm, top) and $D_h(x)$ for charge $-1/3$ quarks (down, strange, bottom), and defining valence distributions as the difference between quark and antiquark distributions, e.g., $U_h^v(x) = U_h(x) - \bar{U}_h(x)$, the structure functions can be directly related to the PDFs:

$$F_{2+}^{(Wh)} = x\sum_D [D_h(x) + \bar{D}_h(x)] + x\sum_U [U_h(x) + \bar{U}_h(x)]; \tag{1.21}$$

$$F_{3+}^{(Wh)} = \sum_D D_h^v(x) + \sum_U U_h^v(x). \tag{1.22}$$

And for the differences:

$$F_{3-}^{(Wh)} = x\sum_D [D_h(x) + \bar{D}_h(x)] - x\sum_U [U_h(x) + \bar{U}_h(x)]; \tag{1.23}$$

$$F_{2-}^{(Wh)} = \sum_D D_h^v(x) - \sum_U U_h^v(x). \tag{1.24}$$

For spin-1/2 partons, the Callan-Gross relation holds:

$$2xF_{1\pm}^{(Wh)} = F_{2\pm}^{(Wh)}. \tag{1.25}$$

Thus, by measuring these structure functions in DIS experiments, it is possible to

extract the parton distribution functions.

1.3 Beyond the Leading Order

Perturbative Corrections

The leading oreder approximation, which only considers the elastic scattering of a lepton off a single parton, is a valuable but ultimately incomplete picture. To obtain accurate and reliable results, it is necessary to include higher-order perturbative corrections that account for more complex processes. In modern analyses, these corrections are typically calculated up to Next-to-Next-to-Leading Order (NNLO). Here, we will provide a brief overview of how Next-to-Leading Order (NLO) corrections are handled for DIS.

Before delving into the details, it is important to introduce the concept of Infrared (IR) safety. In perturbative calculations, certain quantities can exhibit divergences due to soft (low-energy) gluon emission or collinear (parallel) particle emission. However, the Kinoshita-Lee-Nauenberg theorem of QED [5, 6] shows that these divergences cancel out when calculating physically observable quantities that are independent of these phenomena. Observables that exhibit this cancellation are called IR safe and can be consistently calculated at every order in perturbation theory. Modern analyses, however, incorporate higher-order effects from perturbative QCD, such as the emission of gluons from quarks or gluon-gluon interactions. These are referred to as Next-to-Leading Order (NLO), Next-to-Next-to-Leading Order (NNLO), and so on.

The perturbative expansion of an observable, such as a structure function, can be written as a series in powers of the strong coupling constant, α_s:

$$F_2(x,Q^2) = F_2(x,Q^2)^{(LO)} + \alpha_s F_2(x,Q^2)^{(NLO)} + \alpha_s^2 F_2(x,Q^2)^{(NNLO)} + \dots \quad . \tag{1.26}$$

This expansion is only valid in high-energy regimes where $\alpha_s \ll 1$ (the regime of asymptotic freedom). The procedures for calculating these higher order corrections and their dependence on the energy scale, Q^2, are a crucial part of modern PDF analysis and will be further discussed in the following sections.

The Factorization Theorem

The factorization theorem is a central element of modern PDF analysis. It provides a way to separate the short distance, perturbatively calculable parts of a process from the long distance, non-perturbative aspects of hadron structure. Essentially, it allows the cross-section for a hard scattering process to be expressed as a product of

two distinct components: the parton distribution functions (PDFs), which describe the distribution of partons within the hadron, and the partonic cross-sections, which describe the interaction between the individual partons.

Consequently, the structure functions can be written in a factorized form as Eq. 1.17. By making the dependencies explicit and distinguishing the equations, they are presented as:

$$F_a^{(Vh)}(x,Q^2) = \sum_{i=q,\bar{q},G} \int_x^1 \frac{d\xi}{\xi} C_a^{(Vi)} \left(\frac{x}{\xi}, \frac{Q^2}{\mu^2}, \frac{\mu_f^2}{\mu^2}, \alpha_s(\mu^2) \right) \times \phi_{i/h}(\xi, \mu_f, \mu^2) \quad (a = 1, 3); \tag{1.27}$$

$$F_2^{(Vh)}(x,Q^2) = \sum_{i=q,\bar{q},G} \int_x^1 d\xi C_2^{(Vi)} \left(\frac{x}{\xi}, \frac{Q^2}{\mu^2}, \frac{\mu_f^2}{\mu^2}, \alpha_s(\mu^2) \right) \times \phi_{i/h}(\xi, \mu_f, \mu^2). \tag{1.28}$$

In these equations, the index i sums over all contributing partons (quarks, anti-quarks, and gluons). An additional dependence on the factorization scale, μ_f, is introduced. This scale serves as a theoretical boundary: any physics larger than μ_f^2 contributes to the short distance coefficient functions $C_a^{(Vi)}$, while effects from virtualities smaller than μ_f^2 are absorbed into the long- istance PDFs $\phi_{i/h}$.

The power of the factorization theorem lies in the nature of its two components:

- The coefficient functions $C_a^{(Vi)}$ are calculable in perturbative QCD and are independent of long distance effects. They describe the hard scattering dynamics.
- The parton distribution functions $\phi_{i/h}$ are universal and independent of the specific hard scattering process. They describe the intrinsic structure of the hadron.

By calculating the coefficient functions perturbatively and confronting the factorized equations with experimental data, it is possible to extract the non-perturbative PDFs. These extracted PDFs can then be used to make predictions for other hard scattering processes.

Parton Evolution

As previously mentioned, a method is required to extend the predictions for PDFs to different energy scales, Q^2. The renormalization group equation states that if a parton distribution is measured at a specific scale μ, its behavior can be predicted for another scale μ' (provided that the process remains in the perturbative regime). This concept is known as evolution of the structure functions. The evolution of parton distributions is described by the DGLAP (Dokshitzer-Gribov-Lipatov-Altarelli-Parisi)

equations (where the scale is set to $\mu = \mu_f$):

$$\mu^2 \frac{d}{d\mu^2}\phi_{i/h}(x,,\mu,\mu^2) = \sum_{j=q,\bar{q},G}\int_x^1 \frac{d\xi}{\xi}P_{ij}\left(\frac{x}{\xi},\alpha_s(\mu^2)\right)\phi_{j/h}(\xi,\mu,\mu^2). \tag{1.29}$$

The evolution kernels, $P_{ij}(z)$, represent the probability of finding a parton i inside a parton j with a momentum fraction z. These kernels are given by perturbative expansions starting at order $\mathcal{O}(\alpha_s)$. The DGLAP equations govern the dependence of the PDFs on the factorization scale, allowing them to be calculated at any scale μ^2 once they are known at an initial scale Q_0^2. A detailed description of the kernels and the DGLAP equations can be found in [3].

1.4 Experimental Database for PDF Extraction

While the previous sections focused on DIS experiments, the database used for modern PDF fits, including the one for this thesis, involves a broader range of experiment types. The kinematic coverage data used by the NNPDF collaboration is shown in Fig.1.1, and described in [7].

1.4.1 Deep Inelastic Scattering (DIS)

As described earlier, DIS involves a high-energy lepton scattering off an hadron. It can be categorized into two main types:

- Charged Lepton DIS: These experiments, using electrons or muons (e.g., SLAC ,HERA, JLab), are particularly sensitive to the distributions of valence quarks (up and down) and the sea quarks.
- Neutrino DIS: These experiments use neutrinos and antineutrinos, which interact via the weak force. This allows for the separation of quark and antiquark distributions, as the weak interaction distinguishes between them (e.g., NuTev, NOMAD).

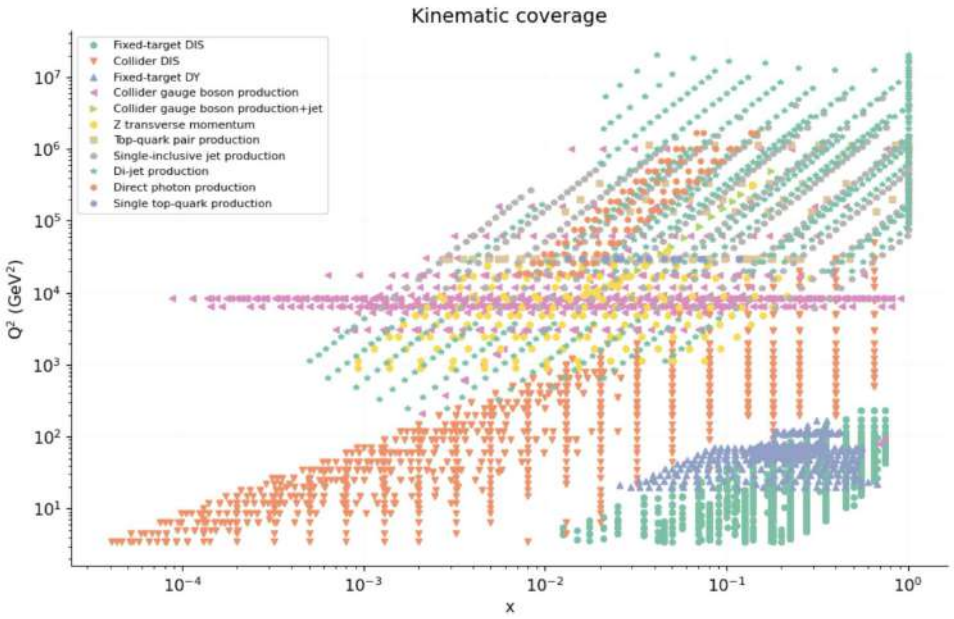

Figure 1.1: Kinematic coverage of the NNPDF collaboration

1.4.2 Hadron Collisions

In hadron-hadron collision experiments, such as those performed at the LHC (ATLAS, CMS, LHCb), two protons are accelerated and collided at high energies. These processes provide direct access to the partonic structure of the proton, in particular to the gluon distribution and linear combinations of quarks-antiquarks PDF, which cannot be fully constrained from deep inelastic scattering (DIS) alone.

The main classes of processes providing complementary information about the proton structure are:

- **Jet Production:** High-energy jets originating from quark–gluon and gluon–gluon scatterings offer a direct probe of the gluon content of the proton.
- **Drell–Yan Process:** The annihilation of a quark and an antiquark into a virtual photon or Z boson, subsequently decaying into a lepton pair (e^+e^-, $\mu^+\mu^-$), provides precise constraints on the quark and antiquark distributions.
- **Top-Quark Pair Production:** $t\bar{t}$ pairs are mainly produced via gluon–gluon fusion at LHC energies, providing a sensitive probe of the gluon PDF at high scales ($Q^2 \sim m_t^2$).

Combining data from these complementary processes is essential to obtain a comprehensive and accurate determination of the parton distribution functions.

Chapter 2

Analysis of Parton Distribution Functions (PDFs)

This chapter introduces the central subject of this thesis: the analysis of parton distribution functions. The discussion begins with a description of the two principal methodologies employed to extract PDFs from experimental data, namely the Monte Carlo method and the Hessian approach, with particular attention devoted to the conversion of results between the two frameworks. Subsequently, the concept of the closure test is presented, together with an explanation of how it is implemented within the NNPDF collaboration. Finally, the problem addressed in this thesis will be presented.

2.1 The Functional Parametrization Method

The functional parametrization method constitutes the traditional approach to the determination of PDFs. As its name suggests, the strategy consists of assuming a functional form for the PDFs, motivated by theoretical considerations, which depends on a set of free parameters subsequently fitted to experimental data. A typical example of such a parametrization is given by:

$$xf(x, Q_0) = a_0\, x^{a_1}(1-x)^{a_2} \exp\left[a_3 x + a_4 x^2 + a_5\sqrt{x} + a_6 x^7\right], \tag{2.1}$$

where x denotes the momentum fraction carried by the parton, Q_0 represents the reference energy scale, and $\vec{a}$ is the vector of free parameters, which generally differs for each parton flavour. For a detailed discussion of the parameterization models, see the works of the MSHT [8] and CTEQ [9] collaborations.

Once the functional form is chosen, the optimal parameter values are determined

through the minimization of an error function, typically a chi-squared estimator:

$$\chi^2(\vec{a}) = \sum_{i,j}^{N_{\rm dat}} \left(t_i(\vec{a}) - m_i\right) \left(Cov\right)_{ij}^{-1} \left(t_j(\vec{a}) - m_j\right), \tag{2.2}$$

where $t_i(\vec{a})$ represents the theoretical prediction associated with the parameter set $\vec{a}$, m_i corresponds to the experimental measurement, and Cov denotes the covariance matrix of the data. The covariance matrix generalizes the concept of variance and covariance to the case of multiple correlated measurements, thereby encoding both statistical and systematic uncertainties, as well as their correlations. The parameter set $\vec{a}_0$ that minimizes the chi-squared function is identified as the best-fit solution.

Error Estimation: The Hessian Method

The Hessian method provides a systematic approach for propagating uncertainties within the functional parametrization framework. The procedure relies on evaluating the variation of the χ^2 function in the vicinity of the best-fit parameters $\vec{a}_0$. In this region, it is assumed that the χ^2 surface can be approximated by a quadratic expansion, namely

$$\Delta\chi^2(\vec{a}) = \sum_{i,j} (a - a_0)_i \, H_{ij} \, (a - a_0)_j, \tag{2.3}$$

where H_{ij} denotes the components of the Hessian matrix,

$$H_{ij} = \frac{1}{2} \left. \frac{\partial^2 \chi^2}{\partial a_i \, \partial a_j} \right|_{\vec{a}=\vec{a}_0}. \tag{2.4}$$

Introducing the displacement vector $\vec{\delta} = \vec{a} - \vec{a}_0$, the above expression simplifies to:

$$\Delta\chi^2 = \vec{\delta}^T H \vec{\delta}. \tag{2.5}$$

Since H is by construction a real and symmetric matrix, it can be diagonalized in terms of its eigenvalues and orthonormal eigenvectors $\{\vec{v}_i\}$, with $i = 1, \ldots, N_{\rm eig}$. Rescaling the eigenvectors by their corresponding eigenvalues λ_i, one obtains a new orthogonal basis $\{\vec{e}_i\}$ such that

$$\Delta\chi^2 = \sum_i z_i^2, \qquad z_i = \vec{\delta} \cdot \vec{e}_i. \tag{2.6}$$

Geometrically, this relation describes a hypersphere of radius $\sqrt{\Delta\chi^2}$ in parameter space. The condition $\Delta\chi^2 = 1$ corresponds to the one-standard deviation confidence region (the 68% confidence level), which is often referred to as the parameter fitting criterion.

Furthermore, we define a covariance matrix in the parameter space, derived from the Hessian matrix according to the relation:

$$H = C^{-1}. \tag{2.7}$$

The eigenvectors of the covariance matrix are the same obtained from those of the Hessian. Consequently, the distribution of the fitted parameters is described by a multivariate Gaussian probability density function,

$$p(\vec{a}) = \frac{(\det C)^{-1/2}}{\sqrt{(2\pi)^N}} \exp\left[-\tfrac{1}{2}\sum_{i,j}^{N}(a-a_0)_i\, C_{ij}^{-1}\,(a-a_0)_j\right], \tag{2.8}$$

which can be interpreted as the product of Gaussian distributions along each independent eigenvector direction.

In practice, however, the condition $\Delta\chi^2 = 1$ is often found to be overly restrictive, leading to an underestimation of the true uncertainties. To account for this discrepancy, a tolerance parameter T is introduced, such that the confidence region is defined by

$$\sqrt{\Delta\chi^2} = T. \tag{2.9}$$

This prescription implies that parameter displacements are permitted to increase the χ^2 function by as much as T^2. The introduction of such a tolerance is motivated by the presence of tensions among different experimental datasets, which may arise from unaccounted systematic uncertainties, theoretical biases such as limitations of the chosen parametrization, or other sources of inconsistency. The value of T is sometimes dependent on the specific eigenvector (dynamic tolerance). Typical values of the tolerance are approximately $T \sim 5$.

2.2 Determination via Neural Networks

The approach adopted by the NNPDF collaboration to extract parton distribution functions (PDFs) is based on machine learning techniques, specifically through the use of neural networks (for a dedicated description see [2]).

Neural networks can be regarded as flexible, parametrized functions defined by a set of free parameters. The basic computational unit of a neural network is the neuron, which transforms an input into an output through the application of an activation function, dependent on a set of adjustable weights and biases. For an input vector x, the action of a single neuron can be expressed as

$$\text{Neuron: } x \mapsto y = f_{\text{activation}}\left(W^T x + \theta\right), \tag{2.10}$$

where W denotes the vector of weights, θ is the bias parameter, and $f_{\text{activation}}$ is a non-linear activation function, which ensures that the network can approximate complex functional forms.

In a neural network, multiple neurons are interconnected. NNPDF uses a lead-forward Network in wich neurons are connected in layers and the inputs to each layer consist on the output of the previous layer. For illustrative purposes, consider a simple feed-forward network composed of three neurons arranged in a 1-2-1 configuration (one input, two hidden neurons, and one output). For an input x, the network output can be written schematically as

$$y = g(\omega_{21}\, g(\omega_{11}x + \theta_1) + \omega_{22}\, g(\omega_{12}x + \theta_2) + \theta_3)\,, \tag{2.11}$$

where ω_{ij} denote the weights, θ_i the biases, and $g(x)$ the activation functions. This example illustrates how even a small network can construct non-linear mappings of the input space.

The specific architecture of the evolution base, obtained from a hyperoptimization process, which is used by the NNPDF collaboration, is represented in 2.1.

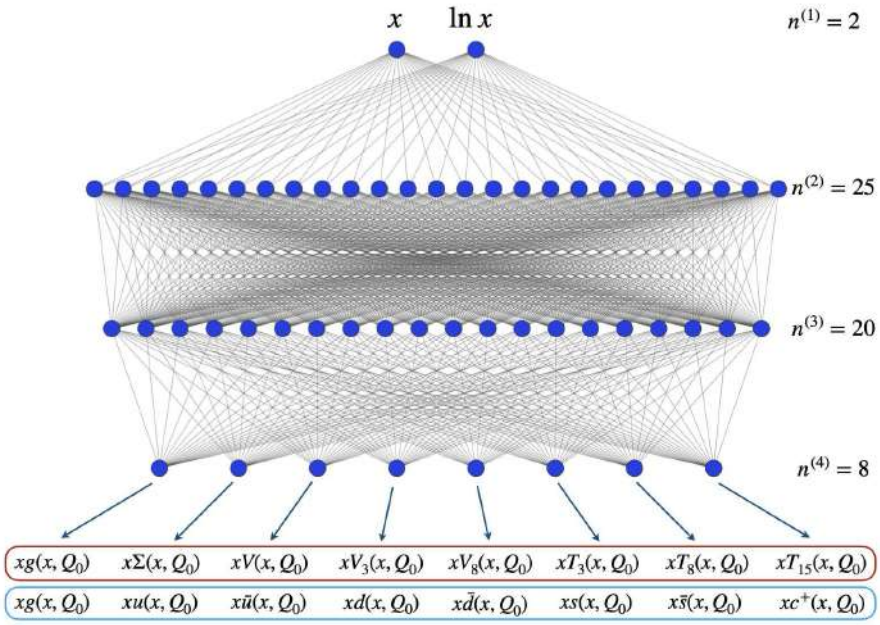

Figure 2.1: *NNPDF network scheme.* For a full description se [2]

Once the network architecture is fixed, the next step is training. During training, the network is exposed to a dataset consisting of input–output pairs, with the aim of adjusting the weights and biases so that the predicted outputs approximate the target values as closely as possible. This optimization is performed by minimizing a loss function, typically the mean squared error as Eq. 2.2. The minimization is commonly performed using gradient descent algorithms, which iteratively update

the parameters in the direction opposite to the gradient of the loss function.

The learning process is subject to several well known challenges, notably underfitting (insufficiently complex networks that fail to capture the structure of the data) and overfitting (overly complex networks that model noise rather than signal). Techniques such as cross-validation, early stopping, and regularization are frequently employed to mitigate these issues.

Once the network architecture is determined and the model successfully trained, the resulting neural networks define PDFs that are not constrained by rigid functional forms. This flexibility allows for a more unbiased and data-driven determination of PDFs compared to traditional parametrizations.

Uncertainty Estimation: The Monte Carlo Method

In contrast to the Hessian method, where uncertainties are assumed to follow a Gaussian distribution around the minimum of the χ^2, the Monte Carlo (MC) approach does not require any prior assumption about the statistical nature of the errors. Instead, it directly estimates the probability distribution of the PDFs through the generation of data replicas.

Starting from the original experimental dataset, one generates an ensemble of N_{dat} pseudo-data replicas, each point drawn according to the probability distribution defined by the input measurements and their covariance. For each replica dataset, an independent fit is performed, minimizing the loss function with respect to the neural network parameters. In the specific case of NNPDF, the loss function is defined in Eq:2.2. Theoretical uncertainties can also be incorporated into this framework (see, e.g., [10]), thereby allowing for a more comprehensive error analysis.

The outcome of this procedure is a set of N_{rep} fitted PDFs, each corresponding to one pseudo-data replica. These PDF replicas collectively represent the probability distribution of the true PDFs given the available data. Statistical estimators for any observable X can then be obtained as the mean and standard deviation over the ensemble of replicas:

$$\langle X \rangle = \frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} X_k, \tag{2.12}$$

$$\sigma^2[X] = \frac{1}{N_{\text{rep}} - 1} \sum_{k=1}^{N_{\text{rep}}} \left(X_k - \langle X \rangle \right)^2, \tag{2.13}$$

where X_k is the value of the observable computed with the k-th replica (for instance, the PDF itself at given (x, Q)).

A major advantage of the Monte Carlo approach is its ability to accurately capture non-Gaussian features of the PDF uncertainties, which are particularly relevant

at small or large values of x. A practical test of Gaussianity is to compare the one standard deviation band with the 68% confidence interval: discrepancies between the two indicate significant departures from Gaussian behaviour.

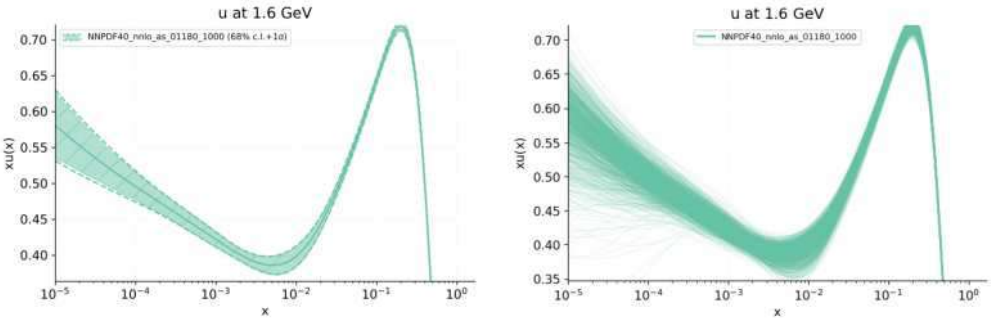

Figure 2.2: *Parton distribution function (PDF) of the u quark at the scale* $Q = 1.6$ *GeV, obtained from a Monte Carlo set of 1000 NNLO replicas with the strong coupling constant* $\alpha_s = 0.118$. *On the right, the plots of all replicas are shown, distributed around the central value; on the left, the central value is displayed together with the standard deviation (dashed line) and the* 68% *confidence interval.*

Multiplicative Uncertainties

In the determination of PDFs, the treatment of experimental systematic uncertainties plays a crucial role in ensuring a consistent and unbiased extraction of the underlying distributions. Among these, multiplicative uncertainties, which scale with the central value of the measured observable, require special care, as their naive inclusion in the covariance matrix can induce artificial biases in the fit.

Experimental uncertainties can generally be classified into three main categories: uncorrelated (statistical) uncertainties σ_i^{uncorr}, correlated additive systematic uncertainties $\sigma_{i,k}^{\text{add}}$, and correlated multiplicative systematic uncertainties $\sigma_{i,k}^{\text{mul}}$. While additive systematics contribute linearly and independently of the observable's magnitude, multiplicative ones (such as luminosity or normalization errors) scale proportionally to the data or theoretical predictions, and therefore influence the overall normalization of the dataset.

A direct use of the experimental central values D_i when constructing the covariance matrix may lead to biased estimations of normalization factors, since the data themselves fluctuate statistically. To address this issue, the NNPDF collaboration introduced the so called t_0-prescription (see [11]), which defines the covariance matrix in terms of a fixed theoretical prediction $T_i^{(0)}$ instead of the fluctuating data values. The corresponding covariance matrix reads:

$$(\mathrm{cov}_{t_0})_{ij} = \delta_{ij}\,(\sigma_i^{\mathrm{uncorr}})^2 + \sum_{k=1}^{N_{\mathrm{add}}} \sigma_{i,k}^{\mathrm{add}}\,\sigma_{j,k}^{\mathrm{add}} + \left(\sum_{k=1}^{N_{\mathrm{mul}}} \sigma_{i,k}^{\mathrm{mul}}\,\sigma_{j,k}^{\mathrm{mul}}\right) T_i^{(0)} T_j^{(0)}, \tag{2.14}$$

where $T_i^{(0)}$ denotes the theoretical prediction corresponding to the data point i. This formulation ensures that multiplicative systematics are consistently treated as relative uncertainties with respect to a stable theoretical reference rather than fluctuating experimental data, thereby avoiding distortions of the fit. In practical implementations, the t_0 values are iteratively updated until self-consistency between the theoretical predictions and the fitted PDFs is achieved.

2.3 Hessian Conversion: The SVD+PCA Method

While both the Hessian and Monte Carlo approaches aim at providing a faithful description of proton structure, they employ different strategies. The Hessian approach is based on the assumption of multivariate Gaussian uncertainties in the parameter space and derives error eigenvectors from the covariance matrix, while The Monte Carlo approach directly propagates the total uncertainties (including both experimental and methodological sources) without relying on Gaussian assumptions. In situations where the Gaussian approximation is valid, it is possible to convert a Monte Carlo set into a Hessian representation, as implemented in the `mc2hessian` code, described in [12].

The conversion proceeds by constructing a Gaussian covariance matrix in PDF space by sampling the PDF grid points x_i from the Monte Carlo replicas. The eigenvectors of this matrix define the corresponding Hessian error directions. The gaussian is centered at the mean of the original Monte Carlo ensemble.

The starting point is the rectangular matrix X of dimensions $N_x N_f \times N_{\mathrm{rep}}$, defined as the difference between each replica $f_\alpha^{(k)}(x_i, Q)$ and the ensemble average (central value) $f_\alpha^{(0)}(x_i, Q)$:

$$X_{lk}(Q) := f_\alpha^{(k)}(x_i, Q) - f_\alpha^{(0)}(x_i, Q). \tag{2.15}$$

Where α labels the N_f independent flavours, i the grid points in x, and k the replica index. The PDF covariance matrix is then given by

$$\mathrm{cov}(Q) = \frac{1}{N_{\mathrm{rep}} - 1} XX^T. \tag{2.16}$$

To analyse this covariance matrix, one performs a singular value decomposition

(SVD) of X:

$$X = U\Sigma V^T. \tag{2.17}$$

Where U and V are orthogonal matrices and Σ is diagonal with the singular values of X. From this decomposition it follows that

$$XX^T = U\Sigma^2 U^T, \tag{2.18}$$

so that the columns of U are the eigenvectors of the covariance matrix, directly corresponding to the Hessian error directions.

In order to reproduce a faithful representation of the PDF, the dimensionality of the eigenvector basis, $N_{\rm eig} = N_x N_f$, is chosen to be large and redundant. An accurate representation is then obtained by retaining only the $\widetilde{N}_{\rm eig} < N_{\rm eig}$ eigenvectors associated with the largest singular values, a procedure equivalent to principal component analysis (PCA). Defining the reduced matrices u, σ, and P, corresponding to the truncated SVD, the Hessian replicas are constructed as

$$\tilde{f}_\alpha^{(k)}(x_i, Q) = f_\alpha^{(0)}(x_i, Q) + \frac{1}{\sqrt{N_{\rm rep} - 1}} (XP)_{lk} \qquad (k = 1, \ldots, \widetilde{N}_{\rm eig}). \tag{2.19}$$

The associated Hessian uncertainties are then obtained as

$$\sigma_{H,\alpha}^{\rm PDF}(x_i, Q) = \sqrt{\sum_{k=1}^{\widetilde{N}_{\rm eig}} \left(\tilde{f}_\alpha^{(k)}(x_i, Q) - f_\alpha^{(0)}(x_i, Q) \right)^2}. \tag{2.20}$$

This method reduces the number of Hessian's eigenvectors needed while retaining the essential statistical information. The resulting compact Hessian representation accurately reproduces the uncertainties of the original Monte Carlo set, while facilitating practical applications such as error propagation in QCD calculations.

2.4 Closure Test

The closure test is a validation procedure whose purpose is to assess whether a given methodology performs correctly for its intended task. Within the NNPDF framework, the closure test was introduced (see Ref. [13]) to evaluate the efficiency and accuracy with which neural networks propagate experimental uncertainties into the extracted PDFs.

Experimental data are typically provided in the form of a central value y_0 accompanied by an experimental covariance matrix $C_{\rm exp}$. In practice, this means that

each data point can be represented as

$$y_0 = f + \eta, \tag{2.21}$$

where f denotes the true value of the observable and η represents the experimental fluctuation in its measurement. Statistically, η is distributed according to a multivariate Gaussian with zero mean and covariance matrix C_{exp}.

In a realistic scenario, the true value f is unknown and corresponds to the physical quantity that one aims to measure. The essence of a closure test is to artificially select a known value of f, from which pseudo-data are generated according to Eq. (2.21). Declaring knowledge of f implies full knowledge of the underlying PDFs, which in turn determine all observables derived from them.

The NNPDF collaboration generates pseudo-data for closure tests at three different levels:

- **Level 0 data:** These are the central predictions of the PDFs without any statistical noise. In the notation of Eq. (2.21), they correspond directly to the true values f.
- **Level 1 data:** Starting from level 0 predictions, statistical fluctuations are introduced according to the experimental covariance matrix C_{exp}. These correspond to the pseudo-measurements y_0.
- **Level 2 data:** These are obtained by further applying Monte Carlo noise to the level 1 pseudo-data, thereby mimicking the behaviour the same methodology used for actual measurements (see Sec. 2.2).

The defining property of closure-test pseudo-data is that their statistical features are fully controlled: they contain no internal inconsistencies and are exactly consistent with the theoretical model used for their generation. Consequently, a fit performed on closure-test data should, within statistical fluctuations, reproduce the original PDFs employed in their construction, with correctly estimated uncertainties. Deviations from this expectation would reveal potential deficiencies in the fitting methodology or in the treatment of uncertainties.

2.5 Problem Statement

A key advantage of the Hessian representation of Parton Distribution Functions (PDFs) lies in the fact that the orthogonal eigenvectors of the Hessian matrix can be treated as nuisance parameters. This feature makes Hessian sets particularly

suitable for phenomenological applications: since experimental analyses already include a wide range of nuisance parameters (such as detector calibration uncertainties or beam luminosity errors), the eigenvector directions of a Hessian PDF set can be consistently incorporated into the propagation of experimental and theoretical uncertainties. Moreover, this framework enables one to identify which eigenvector directions contribute most significantly to the total uncertainty.

In the Hessian approach, the quantity used to evaluate how the quality of the fit varies when moving along a direction in parameter space is the $\Delta\chi^2$. This variable measures the normalized distance (in units of statistical uncertainty) from the central fit. Under ideal statistical assumptions, the expected value of $\Delta\chi^2$ is unity, reflecting the fact that deviations are measured in units of one standard deviation.

However, in practice it is often necessary to introduce a tolerance parameter, T, such that a one–standard–deviation contour corresponds to $\sqrt{\Delta\chi^2} = T$. The presence of such a tolerance accounts for possible inconsistencies or tensions among the fitted data sets, as well as for limitations in the theoretical modeling or experimental uncertainty estimates.

In the NNPDF approach, the one–sigma contour is determined by the standard deviation of the ensemble of Monte Carlo replicas. When the Monte Carlo ensemble is mapped onto a Hessian representation by an Hessian conversion, this contour can be approximated by a multivariate Gaussian distribution. As illustrated in Fig. 2.2, this quite accurate since the one–sigma contour is approximately coincident with the 68% confidence region of the replica sample.

At this stage, it becomes particularly interesting to investigate the behavior of the $\Delta\chi^2$ distribution, as defined in Eq. 2.3, when a Hessian conversion is applied to a Monte Carlo set. This allows one to assess whether it is also necessary to introduce a tolerance T in this context. If such a tolerance were required, it could point to potential inconsistencies within the data. In such case, this hypothesis could be tested by performing the same analysis on the corresponding closure–test set: in which, one would expect $\Delta\chi^2 = 1$ thanks to the internal consistency of the procedure.

Recent studies [14] have reported the need for a tolerance of $T = 1.3 \pm 0.1$; however, these results were obtained using an earlier version of the NNPDF methodology, which has since been updated and may lead to different conclusions in more recent analyses.

Chapter 3

Results

In this chapter the analyses and results obtained are presented. In the first section, the properties of the $\Delta\chi^2$ distribution and the differences between true data and closure test ensembles are discussed. Subsequently, the dependence of the results on the number of replicas is analyzed, considering both quadratic and quartic fits of $\Delta\chi^2$. Finally, the dependence of the results on the number of eigenvectors used in the Hessian conversion is examined.

3.1 Distribution of $\Delta\chi^2$

The first issue addressed concerns the behavior of the $\Delta\chi^2$ distribution in the context of a closure test. For the construction of the closure test ensembles, the reference truth was chosen to be the PDF set `NNPDF40_nnlo_as_0118_1000`, generated using neural network techniques on the full NNPDF collaboration database, starting from a set of 1000 replicas.

As an initial comparison, results are reported for two PDF sets constructed on the full database, where the Hessian conversion was performed using a basis of 100 eigenvectors. This choice follows previous studies in which this number was identified as sufficient for good accuracy [12].

From Fig. 3.1, it can be observed that both distributions are centered around $\Delta\chi^2 \approx 1$, in agreement with theoretical expectations. However, a large spread of $\Delta\chi^2$ is observed, including negative values. This occurred in both the closure test and the true data.

$\Delta\chi^2$ values differing to 1 could be interpreted either as numerical inefficiencies in the minimization procedure of χ^2, which unavoidably lead to some large entries, or as consequences of incompatibilities between different datasets. In the latter case, noise and inconsistencies deform the parabolic structure of the χ^2 surface, resulting in values different from 1. However, since the same behavior is observed

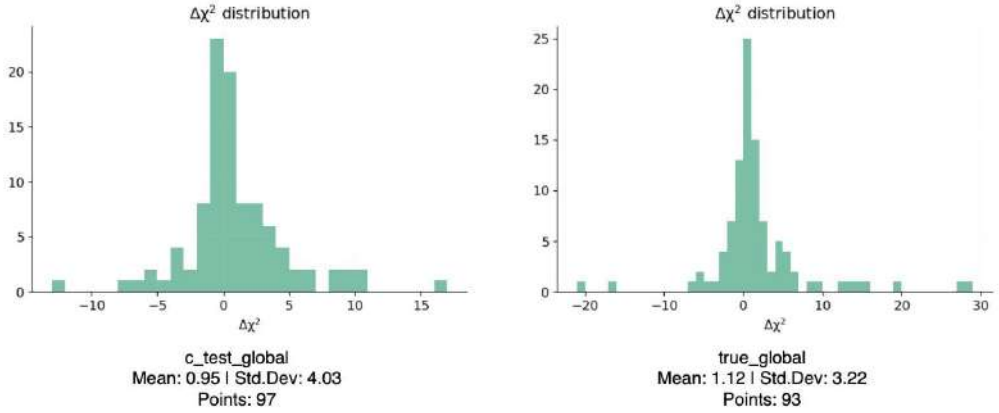

Figure 3.1: *Distribution of $\Delta\chi^2$ for Hessian sets. Left: closure test on the full database. Right: true data. Mean values and standard deviations are computed using entries within two standard deviations from the central value.*

in the closure test, where such inconsistencies cannot arise by construction, the second hypothesis can be excluded. In order to investigate the contribution of the minimization inefficencies the DIS-only subspace of the database was studied. In this case, the data space is reduced to consistent measurements originating from the same type of experiments, making minimization inefficiencies less likely. The results are shown in Fig. 3.2.

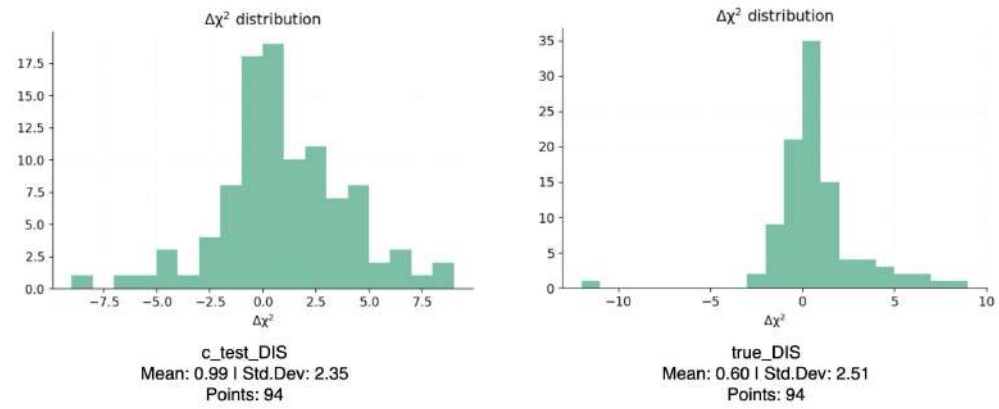

Figure 3.2: *Distribution of $\Delta\chi^2$ for Hessian sets in the DIS-only subspace. Left: closure test results. Right: true data.*

Also in this case, a large spread of $\Delta\chi^2$ value is observed, since both the closure test and the true data display similar entries, as shown in 3.3.

The fact that the results of the closure test exclude the possibility that the dispersion of the $\Delta\chi^2$ values can be attributed to inconsistencies among the data, and that those obtained in the DIS subspace rule out potential inefficiencies in the minimization procedure, leads to the conclusion that Neural Network method does

not rely solely on the goodness-of-fit (χ^2) as the likelihood criterion for a given PDF configuration. Nevertheless, it should be noted that, although this phenomenon occurs, the distribution of values still appears to be centered around unity, exhibiting a behavior that warrants further investigation.

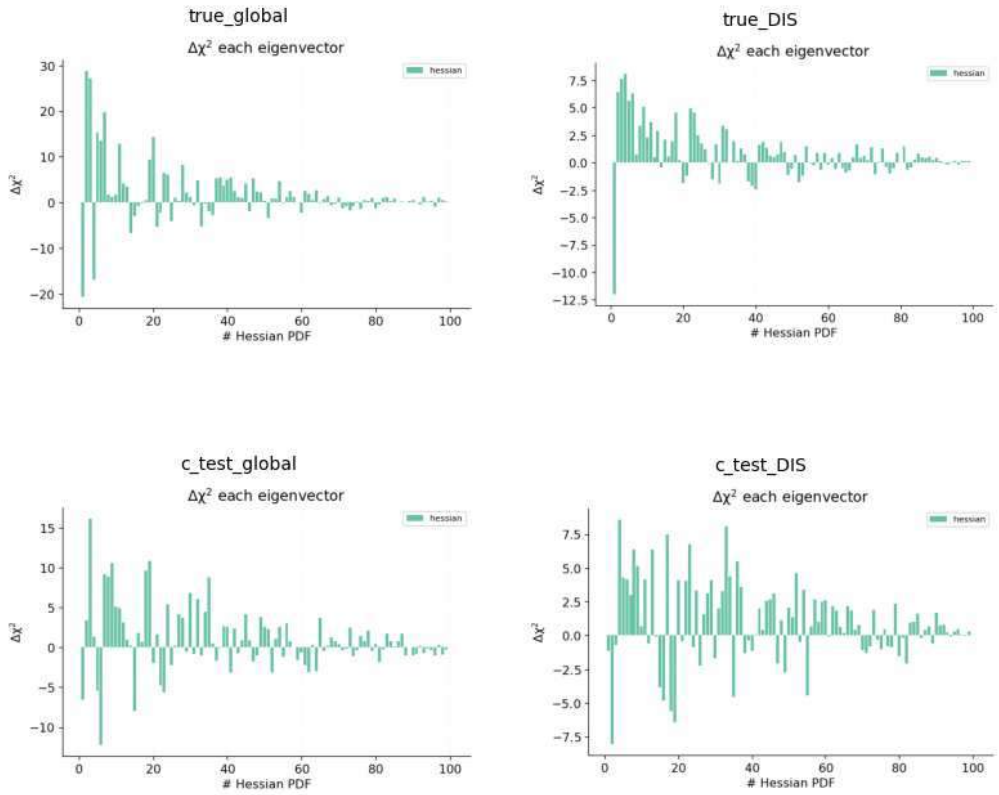

Figure 3.3: *Distribution of the single $\Delta\chi^2$ values for each eigenvector. Top: true data. Bottom: closure test*

An additional aspect that emerges from this comparison is the discrepancy between the DIS-only subset and the full dataset in the case of true data. While the closure test maintains a mean value close to unity in both cases, the distribution for the true data falls below unity when the full dataset is considered. This behavior, will be further investigated in the following sections.

3.2 $\Delta\chi^2$ Shape and Fitting Methodologies

At this stage, it was investigated whether the results obtained in the previous section depend on the shape of the $\Delta\chi^2$ function around its central value or on the size of the Monte Carlo sample employed in the analysis.

To perform this study, the $\Delta\chi^2$ can be expanded along each direction as a function of the displacement $\theta - \theta_0$ around the central value:

$$\Delta\chi^2(\theta) = \chi^2(\theta) - \chi^2(\theta_0) \tag{3.1}$$

$$= a\frac{(\theta-\theta_0)}{\sigma} + b\left[\frac{(\theta-\theta_0)}{\sigma}\right]^2 + c\left[\frac{(\theta-\theta_0)}{\sigma}\right]^3 + d\left[\frac{(\theta-\theta_0)}{\sigma}\right]^4, \tag{3.2}$$

where the coefficient a quantifies possible fitting inefficiencies, b is related to the tolerance parameter, and c and d account for non-parabolic deviations from the quadratic approximation. The shape analysis was carried out by fitting both the general quartic form in Eq. (3.2) and a simpler quadratic form where $a = c = d = 0$.

The $\Delta\chi^2$ values were computed as 2-sigma clipped averages across the $\Delta\chi^2$ values of each eigenvector for the set of displacements $(\theta - \theta_0) = \frac{1}{k} = \{0.125, 0.25, 1, 2, 4\}$, allowing us to probe the local shape of the $\Delta\chi^2$ function around its minimum.

Furthermore, the analysis was repeated for Monte Carlo ensembles with different numbers of replicas, $N_{\text{reps}} = \{150, 200, 250, 300, 350, 400\}$, in order to assess whether the results depend on the sample size. The fits were performed both for the closure test and for the true data, and the results for the quadratic and the quartic fit are shown in Figs. 3.4 and 3.5.

The trends of the fitted coefficients, together with their uncertainties as a function of the number of replicas, are reported in Fig. 3.6.

From the plots, it can be observed that in the quartic expansion the coefficients $\{a, c, d\}$ remain significantly below unity, while the parameter b shows stable behavior across all replica sets. Moreover, the uncertainties obtained from the quadratic fit are smaller and more consistent than those from the quartic fit, suggesting that the quadratic approximation provides a more reliable description of the parameter trend.

Furthermore, in contrast to the findings in [15], which used an older derivation methodology for the PDFs, the results obtained with the latest version (in its final form) also indicate no significant dependence of the parameter b on the number of replicas, as its variations are well explained by statistical fluctuations. These considerations lead to the conclusion that b represents a suitable and robust choice for quantifying the tolerance of the distributions.

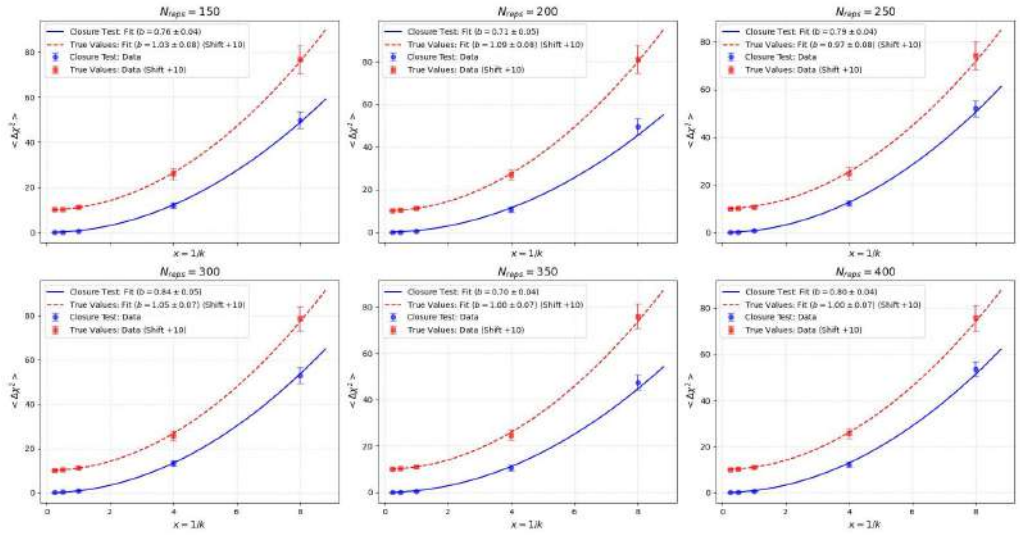

Figure 3.4: *Quadratic fit of the averaged $\Delta\chi^2$ for the closure test (blue) and true data (red) as a function of the number of replicas. For visualization purposes, the true data are shifted upward by 10 units.*

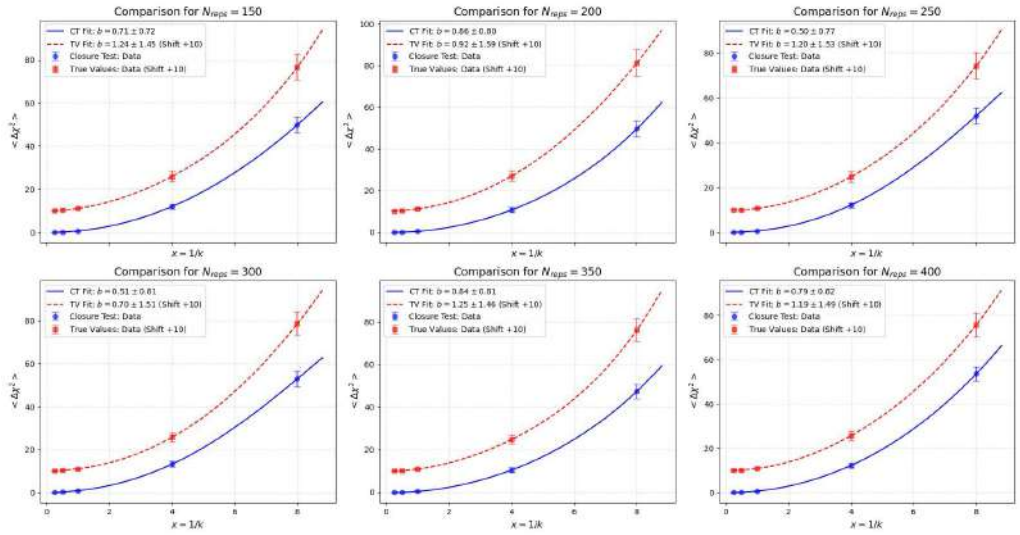

Figure 3.5: *Quartic fit of the averaged $\Delta\chi^2$ for the closure test (blue) and true data (red) as a function of the number of replicas. For visualization purposes, the true data are shifted upward by 10 units.*

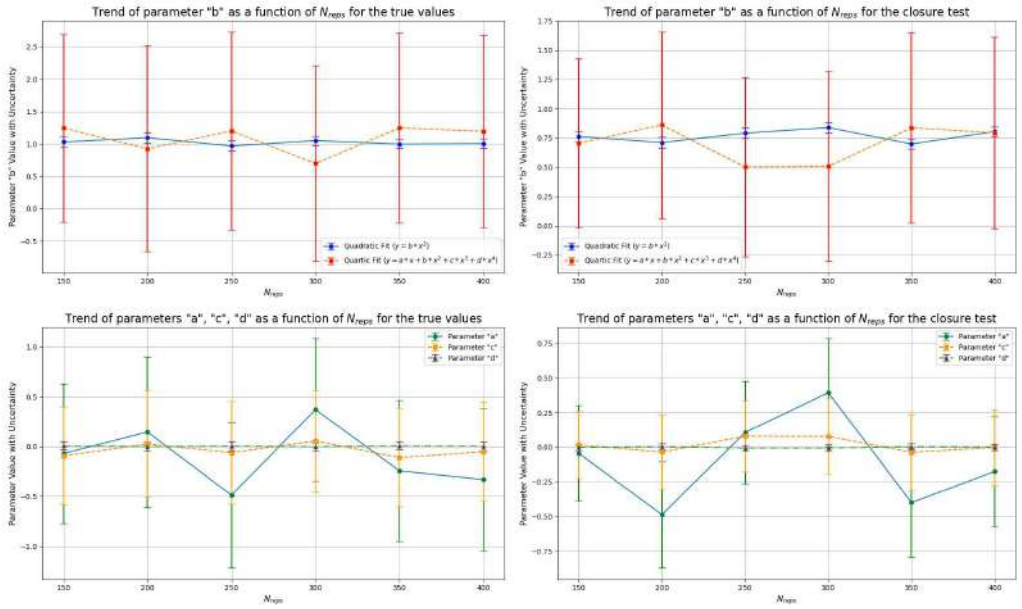

Figure 3.6: *Comparison of fitted parameters: quadratic coefficient b (top) for true data (left) and closure test (right); quartic coefficients a, c, and d (bottom).*

Fit Scenario	**Fit Function**	**Mean Parameter** $\bar{b} \pm \sigma_{\bar{b}}$
True Datas	Quadratic	1.0237 ± 0.0309
	Quartic	1.0836 ± 0.6146
Closure Test	Quadratic	0.7675 ± 0.0177
	Quartic	0.7011 ± 0.3221

Table 3.1: *Mean values with their uncertainties (calculated as the standard deviation of the mean) of the parameter b obtained from quadratic and quartic fits applied to the true data and to the closure test after the Hessian conversion with* 100 *eigenvectors.*

Finally, it can be noted (see Table 3.1) that the mean value obtained for the true data is $b_{\text{true}} \sim 1$, whereas the closure test yields $b_{\text{closure}} < 1$. Before concluding that a genuine discrepancy exists between the two cases, it should be considered that these results were derived from a Hessian conversion based on 100 eigenvectors, chosen as optimal for representing the true data. However, this assumption may not hold for the closure test. Therefore, the dependence of the results on the number of eigenvectors will be examined in the following section.

3.3 Dependence on the Number of Eigenvectors

As discussed previously, this section focuses on the study of the $\Delta\chi^2$ behavior as a function of the number of eigenvectors selected for the Hessian conversion. The analysis was performed on the same datasets used in the previous section, i.e., $N_{\text{reps}} = \{150, 200, 250, 300, 350, 400\}$, for both the closure test and the true data. Once the Hessian conversion is carried out for a given N_{eig}, the eigenvectors are ordered in decreasing order according to their corresponding eigenvalues, and the moving average of the $\Delta\chi^2$ values is computed. The results of this analysis are presented in Figs. 3.8 and 3.9.

In the case of the true data (Fig. 3.8), a plateau region can be observed around $\Delta\chi^2 = 1$ for $80 < N_{\text{eig}} < 100$. For larger values of N_{eig}, the mean $\Delta\chi^2$ tends toward zero. This behavior indicates that, for the smaller eigenvalues, a variation of θ around θ_0 produces almost no change in the PDFs. Consequently, the contribution to $\Delta\chi^2$ from these modes vanishes, driving the moving average toward zero. The dependence of $\Delta\chi^2$ on the eigenvalue magnitude is clearly illustrated in Fig. 3.10, where the $\Delta\chi^2$ values are arranged in decreasing order of eigenvalue, with the dominant contributions corresponding to the largest eigenvalues.

For the closure test (Fig. 3.9), the plateau region is instead observed for $40 < N_{\text{eig}} < 60$. At $N_{\text{eig}} = 100$, the system already enters an overfitting regime, where the mean $\Delta\chi^2$ falls below unity.

A direct comparison between the two cases, showing the full trends across all replica sets, is provided in Fig. 3.7.

This behavior indicates that the true data are inherently noisier than the closure test, requiring a larger number of eigenvectors to properly represent the underlying uncertainties. However, this does not point to any fundamental issue with the method itself, as the results remain statistically compatible once the appropriate eigenvectors range is chosen.

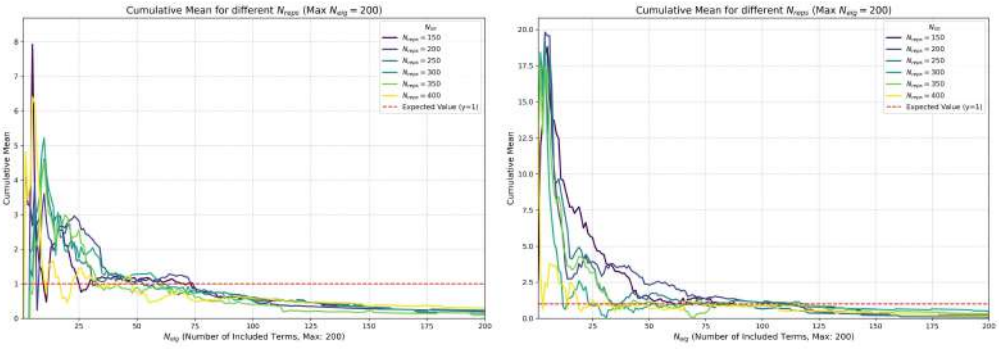

Figure 3.7: *Comparison between the moving averages of $\Delta\chi^2$ values for the closure test (left) and the true data (right) across the different replica sets.*

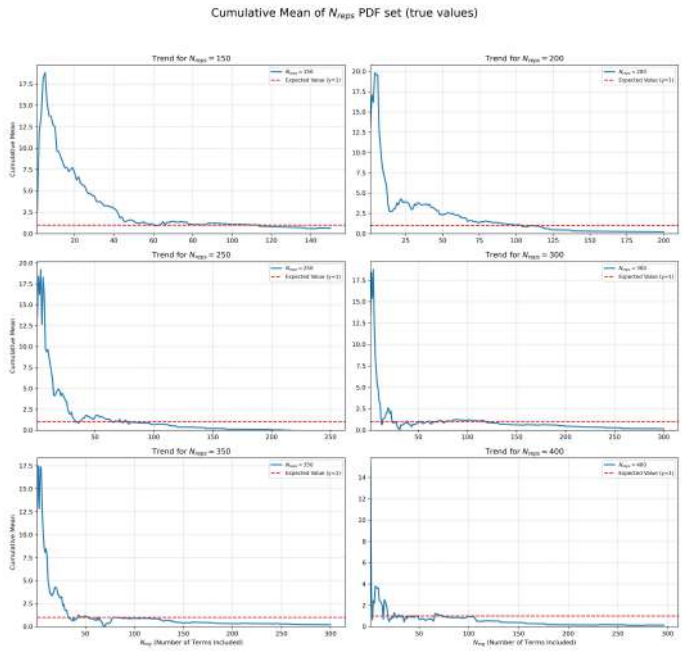

Figure 3.8: *Mean value of $\Delta\chi^2$ as a function of the number of eigenvectors for datasets with different numbers of replicas, computed on the true data. The red line indicates the reference value $\Delta\chi^2 = 1$ (note the different scales).*

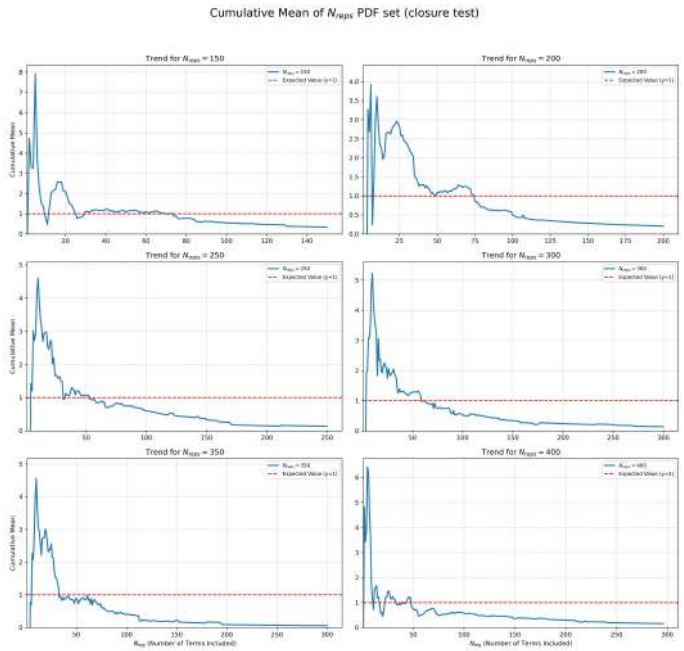

Figure 3.9: *Mean value of $\Delta\chi^2$ as a function of the number of eigenvectors for datasets with different numbers of replicas, computed on the closure test. The red line indicates the reference value $\Delta\chi^2 = 1$ (note the different scales).*

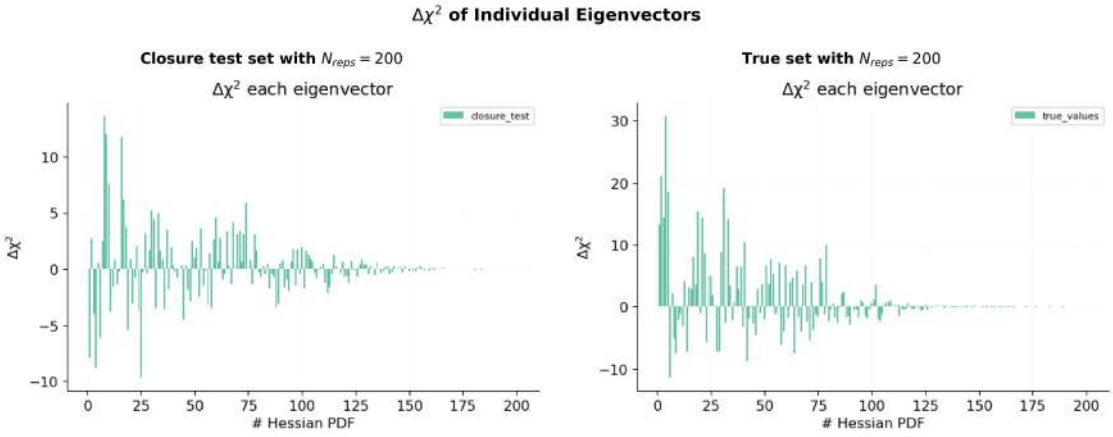

Figure 3.10: *Values of $\Delta\chi^2$ computed for each individual eigenvector, ordered in decreasing order of their eigenvalues. The plots clearly show that the dominant contributions arise from the eigenvectors associated with the largest eigenvalues.*

As a further confirmation of the previous observation, a quadratic fit was repeated, following the procedure described in the previous section, for the Hessian conversion onto a basis of 50 eigenvectors in the closure test case. The evolution of the fitted parameter b is shown in Fig. 3.11.

With this configuration, the same results as those obtained for the true data are recovered, yielding a mean value of:

$$b_{closure} = 1.0322 \pm 0.0019. \tag{3.3}$$

This confirms that, for the closure test, a Hessian conversion based on approximately 50 eigenvectors is sufficient to provide a stable and accurate description of the system.

Furthermore, this analysis allows us to conclude that no inconsistencies are observed between the shape of the $\Delta\chi^2$ obtained from the true data and that from the closure test, thereby validating the analysis presented in Section 3.1.

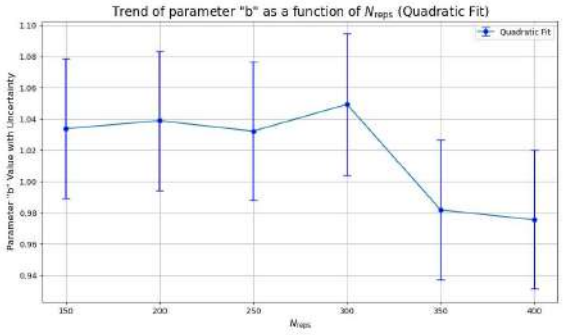

Figure 3.11: *Mean values of the parameter b obtained from the quadratic fit, with their corresponding uncertainties, for the Hessian conversion of the closure test using 50 eigenvectors.*

3.4 Conclusions and Outlook

The results presented in this thesis provide a characterization of the behavior of Monte Carlo PDF sets when analyzed through their Hessian conversion.

An initial investigation of the $\Delta\chi^2$ distribution revealed that individual eigenvector contributions could attain both large and negative values. Since the same behavior is observed in the closure test, however, it can be excluded that such effects arise from inconsistencies among the experimental datasets. Furthermore, by repeating the analysis on a DIS-only data subset, the hypothesis of minimization inefficiencies can also be ruled out.

A second analysis, aimed at studying the shape of the $\Delta\chi^2$ distribution around the central value and its dependence on the number of replicas, demonstrated that no dependence on the number of replicas are present within this methodology and that quartic deviations from the quadratic approximation are negligible.

Finally, since the mean values obtained for the closure test were systematically lower than those derived from the true data, an additional study was conducted to explore the dependence on the number of eigenvectors used in the Hessian conversion. This analysis showed that, while the optimal description for true data is achieved around $N_{\text{eig}} = 100$, the closure test reaches a stable plateau already at $N_{\text{eig}} \approx 50$. This difference reflects the higher level of noise intrinsic to the true data, which therefore require a larger number of eigenvectors to achieve an equivalent representation. Moreover, when performing the quadratic fit on the closure test configuration with $N_{\text{eig}} = 50$, the resulting values are consistent with those obtained for the true data with $N_{\text{eig}} = 100$, confirming the internal consistency of the approach.

An interesting feature emerging from this study is that, despite the large fluctuations observed in the individual $\Delta\chi^2$ values, the central value of the distribution remains stably centered around unity in the optimal configuration. This behavior suggests the presence of an intrinsic statistical mechanism that preserves the expected normalization of the $\Delta\chi^2$ distribution.

Future investigations could aim at a more rigorous determination of the plateau region by employing the explained variance ratio as a diagnostic criterion applied to the moving average of the $\Delta\chi^2$ values as a function of the number of eigenvectors.

In parallel, deviations from Gaussianity in the Monte Carlo ensemble could be quantified through the Kullback–Leibler divergence, thereby providing a means to test the validity of the Gaussian approximation underlying the Hessian conversion.

Moreover, it remains to be understood why, although the $\Delta\chi^2$ values associated with individual eigenvectors display significant fluctuations, their distribution nevertheless remains centered around the mean value $\langle\Delta\chi^2\rangle = 1$.

Bibliography

[1] https://nnpdf.mi.infn.it.

[2] Richard D. Ball, Stefano Carrazza, Juan Cruz-Martinez, Luigi Del Debbio, Stefano Forte, Tommaso Giani, Shayan Iranipour, Zahari Kassabov, Jose I. Latorre, Emanuele R. Nocera, Rosalyn L. Pearson, Juan Rojo, Roy Stegeman, Christopher Schwan, Maria Ubiali, Cameron Voisey, and Michael Wilson. "The Path to Proton Structure at 1% Accuracy: NNPDF Collaboration". *The European Physical Journal C*, 82(5), May 2022.

[3] George Sterman, John Smith, John C. Collins, James Whitmore, Raymond Brock, Joey Huston, Jon Pumplin, Wu-Ki Tung, Hendrik Weerts, Chien-Peng Yuan, Stephen Kuhlmann, Sanjib Mishra, Jorge G. Morfín, Fredrick Olness, Joseph Owens, Jianwei Qiu, and Davison E. Soper. "Handbook of Perturbative QCD". *Reviews of Modern Physics*, 1995.

[4] Michael E. Peskin and Daniel V. Schroeder. *"An Introduction to Quantum Field Theory"*. Westview Press, Boulder, CO, 1995.

[5] Toichiro Kinoshita. "Mass Singularities of Feynman Amplitudes". *J. Math. Phys.*, 3:650, 1962.

[6] T. D. Lee and Michael Nauenberg. "Degenerate Systems and Mass Singularities". *Phys. Rev.*, 133:B1549–B1562, 1964.

[7] Richard D. Ball, Stefano Carrazza, Juan Cruz-Martinez, Luigi Del Debbio, Stefano Forte, Tommaso Giani, Shayan Iranipour, Zahari Kassabov, Jose I. Latorre, Emanuele R. Nocera, Rosalyn L. Pearson, Juan Rojo, Roy Stegeman, Christopher Schwan, Maria Ubiali, Cameron Voisey, and Michael Wilson. "An Open-Source Machine Learning Framework for Global Analyses of Parton Distributions", 2021.

[8] J. McGowan, T. Cridge, L. A. Harland-Lang, and R. S. Thorne. "Approximate N^3LO Parton Distribution Functions with Theoretical Uncertainties: MSHT20aN^3LO PDFs". *The European Physical Journal C*, 83(3):185, 2023.

[9] Tie-Jiun Hou, Jun Gao, T.J. Hobbs, Keping Xie, Sayipjamal Dulat, Marco Guzzi, Joey Huston, Pavel Nadolsky, Jon Pumplin, Carl Schmidt, Ibrahim Sitiwaldi, Daniel Stump, and C.-P. Yuan. "New CTEQ Global Analysis of Quantum Chromodynamics with High-Precision Data from the LHC". *Physical Review D*, 103(1), January 2021.

[10] The NNPDF Collaboration, Rabah Abdul Khalek, Richard D. Ball, Stefano Carrazza, Stefano Forte, Tommaso Giani, Zahari Kassabov, Emanuele R. Nocera, Rosalyn L. Pearson, Juan Rojo, Luca Rottoli, Maria Ubiali, Cameron Voisey, and Michael Wilson. "A First Determination of Parton Distributions with Theoretical Uncertainties", 2019.

[11] Richard D. Ball, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, José I. Latorre, Juan Rojo, and Maria Ubiali. "Fitting Parton Distribution Data with Multiplicative Normalization Uncertainties". *Journal of High Energy Physics*, 2010(5), May 2010.

[12] Stefano Carrazza, Stefano Forte, Zahari Kassabov, José Ignacio Latorre, and Juan Rojo. "An Unbiased Hessian Representation for Monte Carlo PDFs". *The European Physical Journal C*, 75(8), August 2015.

[13] Richard D. Ball, Valerio Bertone, Stefano Carrazza, Christopher S. Deans, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, Nathan P. Hartland, José I. Latorre, Juan Rojo, and Maria Ubiali. "Parton Distributions for the LHC Run II". *Journal of High Energy Physics*, 2015(4), April 2015.

[14] Nicola Lambri. "Optimized Regression Models for Parton Distribution Functions Determination Using Deep Learning Methods". Master's thesis, University of Milan, 2020.

[15] Luca Talon. "Optimization of Parton Density Uncertainties". Master's thesis, University of Milan, 2018.