UNIVERSITA' DEGLI STUDI DI MILANO
Facoltà di Scienze Matematiche, Fisiche e Naturali

Corso di laurea triennale in Fisica

The Impact of Correlated Systematics in a Global PDF Analysis

Relatore: Prof. Stefano Forte
Correlatore: Dott. Juan Rojo Chacon

PACS: 12.38.-t

Alice Maria Donati
matricola 707007

Anno Accademico 2008 - 2009

Abstract

It has been suggested that parton densities global analysis is insensitive to the inclusion of correlated systematics. In this work we discuss this statistical issue by comparing two parton sets: the first is obtained including correlations; the second is produced neglecting correlated uncertainties and simply adding in quadrature all systematics. Both parton sets have been determined, from a set of purely deep-inelastic scattering data, using the NNPDF method, based on a Monte Carlo sampling and a neural network parametrization.

Contents

Introduction

In view of the forthcoming experiments at LHC, the physics of strong interactions has been turned into a precision science: a solid knowledge of the structure of nucleons together with the ability to make accurate predictions about cross sections for deep inelastic scattering are now mandatory.

In this context, the problem of a faithful determination of parton distribution functions has been addressed by many collaborations (CTEQ, Alehkin, MSTW, NNPDF).

In QCD, the cross section for deep inelastic scattering is expressed, thanks to factorization theorem, as the convolution of perturbatively computable parton cross sections, times parton densities: this way, the process is related to the underlying scattering off the subconstituents of the nucleon. However, parton densities cannot be determined from first principles yet, but need to be determined from experimental data.

Various methods have been developed to this purpose. Here we are interested in the NNPDF method: parton densities are parametrized by neural networks, so that the parametrization bias is minimized. In order to give a statistically sound representation of uncertainties, a Monte Carlo approach is employed: data points are replicated so that, by fitting neural pdfs to each replica, a Monte Carlo ensemble of pdfs is obtained and all statistical information about the original data is retained.

In this work, we concentrate on the treatment of correlated systematics. Some groups of pdf determination properly include correlations in their parton global analysis (CTEQ, Alehkin, NNPDF). In the NNPDF approach, information about correlated uncertainties is included both at the level of Monte Carlo generation of replicas and at the level of the fitting procedure. However, other groups believe that results are the same whether correlations are treated properly or if they are just added in quadrature.

The purpose of this work is to clarify this statistical issue. We have produced two parton sets: the first is obtained using the standard settings of NNPDF parton analysis, so that correlations are properly considered. For the second, we have neglected correlations: at the level of Monte Carlo generation, we allow correlated points to fluctuate independently; at the level of minimization, we use a diagonal error function, where systematics are added in quadrature. An analysis of the impact of correlations is then made by comparing the two parton sets.

1 Scattering and the Strong Interactions

The forthcoming experiments at LHC are based on proton-proton collision and therefore require a solid understanding of the proton structure.

The proton is a finite sized object with a substructure made up of strong interacting point-like particles, the quarks and the gluons, generically called partons.
The theoretical framework that explains the physics of quarks and gluons is called quantum chromo-dynamics (QCD); it is a quantum field theory, based on a fundamental symmetry referred to as color charge (hence the name "chromo").

Our knowledge of the nucleon structure comes primarily from experiments of deep inelastic scattering (DIS) with lepton beams and nucleon targets; "deep", because the distances probed are comparable or smaller than those peculiar to the nucleon; "inelastic", because initial and final states are not the same.
The main observable of scattering is the differential cross section, i.e. the probability of finding a particle with a given energy within a fixed angle, after the scattering process. From experimental data we can extract information about the substructure of the target and the interactions taking place during the collision.
The cross section for DIS off nucleons is parametrized by the so-called structure functions, whose shape can be determined using experimental data.

DIS off nucleons is determined by the underlying scattering processes between lepton projectiles and partons. In QCD, structure functions are given by the convolution of a parton distribution function (pdf) and a hard cross section, up to corrections suppressed by powers of the ratio of the nucleon scale to the large energy scale of the scattering process. A parton distribution is a quantity that can be roughly understood as the probability of finding a parton inside the nucleon carrying a certain amount of the nucleon momentum; the hard cross section gives the probability of a scattering off the parton.
This latter quantity can be calculated by means of perturbative QCD. On the contrary, pdfs can not be calculated perturbatively, and thus must be extracted from experiments, for instance by assuming a parametrization and then tuning its parameters so that they fit the experimental data.

1.1 Scattering

Scattering is the general phenomenon where some form of radiation or flux, e.g. a beam of particles, deviates from its straight trajectory by interactions with one or more objects (*target*) on its path.

By measuring the kinematical features of the initial and the final states, we may learn about the properties of the scatterer.
The higher the momentum of the projectile the shorter the distances it probes: when the *De Broglie's wavelength*, $\lambda = h/p$, of the exchanged particle is smaller or comparable to the target's size, the internal structure of the latter becomes visible.

1.1.1 Kinematics

We will consider scattering experiments between relativistic particles, therefore their kinematics is described by 4-vectors.
We call p and E the momentum and the energy of the beam particle, P the target's momentum; final states are labelled with prime letters.
If $q = p - p'$, energy and momentum conservation imply that

$$\begin{aligned} (P')^2 = (P+q)^2 &= M^2 + 2Pq + q^2 &(1)\\ &= M^2 + 2M\nu - Q^2 \quad , &(2)\end{aligned}$$

where the Lorentz invariant quantity $\nu = \frac{Pq}{M}$ is the energy transfer $E - E'$ in the frame of reference where the target is at rest. The positive definite quantity $Q^2 = -q^2$ represents the momentum transfer.

Now elastic and inelastic scattering need be distinguished.

- In *elastic scattering* ($a + b \longrightarrow a' + b'$) particles before and after the collision are identical up to energies and momentum; the invariant mass is unchanged and the energy transfer is only due to recoil, i.e the target changes just its kinetic energy. It is easy to see, using momentum conservation, that if the beam energy and the scattering angle are known the energy and the momentum transfers are fixed too:

$$P'^2 = M^2 = M + 2M\nu - Q^2 \tag{3}$$

$$\Longrightarrow 2M\nu - Q^2 = 0 \quad \Longrightarrow \nu = E - E' = \frac{Q^2}{2M} \quad . \tag{4}$$

- In *inelastic scattering* ($a + b \longrightarrow a' + X$, where X is the generic state resulting from the process) the transferred energy may excite the colliding particles into different internal states, changing their nature, or break them into a number of fragments.
  Let W be the invariant mass of the resulting particle (or particles), then

  $$P'^2 = W^2 = M + 2M\nu - Q^2 \tag{5}$$

  $$W > M \quad \Longrightarrow \quad 2M\nu - Q^2 > 0 \quad . \tag{6}$$

  As a result, inelastic scattering has a further degree of freedom: to determine the momentum transfer Q^2 we need to measure both the scattering angle and the energy E' of the scattered particle.

In the experiments, we are able to set the energy E and the momentum p of the beam particles and to measure the scattering angle θ and the energy E' of the product particles, together with the reaction rate within any fixed direction.

1.1.2 Cross Section

The main observable of scattering experiments is the *differential cross section*, the probability that a particle is scattered through a given angle with a given energy. It is a measurable quantity thanks to the following relation:

$$\dot{N}(E, \theta, \Delta\Omega) = \Phi_{\text{beam}} \frac{d^2\sigma(E, E', \theta)}{d\Omega dE'} \Delta\Omega\Delta E' \tag{7}$$

where $\dot{N}$ is the rate of particles measured by a detector with cone of observation $\Delta\Omega$, collecting particles scattered through θ-direction with energy E'; Φ_{beam} is the flux of beam particles (i.e. the number of particles per unit area and unit time).
The *total* or *integral cross section* is defined as:

$$\sigma_{\text{tot}} = \int_0^{E_{\max}} \int_{4\pi} \frac{d^2\sigma(E, E', \theta)}{d\Omega dE'} \, \mathrm{d}\Omega \, \mathrm{d}E' \tag{8}$$

it is therefore a measure of the effective surface area seen by the beam particles, and as such is expressed in units of area.

In quantum mechanics, the usual approach to scattering consists in considering the particles before and after the collision, i.e. far from the interaction region, as free particles (represented by plane waves). The matrix elements of the time evolution operator $S(T, -T)$ between states in the distant past and distant future provide the amplitude for a transition between them. Therefore the cross section is

$$\mathrm{d}\sigma = \lim_{T\to\infty} \frac{1}{T} \frac{1}{|\vec{J}_{\mathrm{inc}}|} |\langle f|S(T, -T)|i\rangle|^2 \, \mathrm{d}E' \, \mathrm{d}\Omega, \tag{9}$$

where $\Phi_{\mathrm{beam}} = |\vec{J}_{\mathrm{inc}}|$ is the flux of the incoming particles.

The cross section for scattering off a system of particles can be related to the elementary cross section for scattering off individual particles through a suitable function of the kinematic variables.
For elastic scattering, these functions are called *form factors*; they are the Fourier transform of the space distribution of the point scatterers.
For deep inelastic scattering, these functions are called *structure functions*, and depend on two kinematical variables.

1.2 Nucleon Structure

High energy scattering experiments are required in order to probe the structure of nucleons, whose radius is about 0.8 Fm.
Elastic scattering off nucleons, from some hundreds MeV up to several GeV, reveals that nucleons are finite sized objects: in fact, the form factors decrease exponentially with the scale.
Further increasing the energy, the form factors eventually become scale-free to a good approximation, as one would expect of scattering off free point-like constituents. More detailed investigations reveal that these constituents carry the same quantum numbers as quarks, which had been suggested as constituents of the nucleons on the basis of spectroscopic arguments.

The Feynman parton model provided the first theoretical framework to explain DIS: the nucleon is considered as a composite system of point-like objects, the partons, each carrying a fraction of the nucleon's energy and momentum. The scattering process involves individual partons, rather than the nucleon as a whole, i.e. the outcome of DIS is the incoherent sum of elastic scattering off partons. Over the short distances and time scales of the collision, the internal interactions among partons can be neglected, so that the struck parton can be regarded as effectively free.

QCD is now well established as the theory of strong interactions and it has been developed to high accuracy. The parton model, which is consistent with QCD at leading order, provides the basis for an intuitive understanding of the physics of quarks and gluons at large scale.

1.2.1 Structure Functions

The cross section of DIS off nucleons is parametrized by the so-called structure functions. As we have previously discussed, two kinematical variables are required. A common choice for these are the scale Q^2 and the *scaling variable*

$$x := \frac{Q^2}{2M\nu}, \qquad 0 < x \leq 1, \tag{10}$$

which is a measure of the inelasticity of the process (for elastic scattering $x = 1$) as shown by eq. (6).

The scaling variable can be approximately understood as the fraction of nucleon momentum carried by the parton involved in the scattering process. If we assume that the momenta of partons are parallel (collinear) to that of the nucleon, and that the parton initially carries a fraction z of the nucleon's momentum, then $p = zP$; since the lepton-parton scattering is elastic, its final momentum is given by:

$$(zP + q)^2 = 0 \tag{11}$$

$$z^2P^2 + 2Pqz + q^2 = 0 \tag{12}$$

$$\Rightarrow z = \frac{Q^2}{2M\nu} = x. \tag{13}$$

Therefore the scaling variable x coincides with the fraction of the nucleon momentum carried by the struck parton, up to mass corrections.

For DIS off nucleons involving an electromagnetic interaction, two structure functions are required to parametrize the cross section; it can be shown that

$$\frac{d^2\sigma}{dx\,dQ^2} = \frac{\alpha^2}{4E^2\sin^4\frac{\theta}{2}}\Big\{\frac{1}{\nu}F_2(x,Q^2)\cos^2\frac{\theta}{2} + \frac{2}{M}F_1(x,Q^2)\sin^2\frac{\theta}{2}\Big\} \tag{14}$$

where α is the coupling constant of electromagnetic interaction.
F_1 and F_2 are assessed by measuring the cross section at different scattering angles for fixed energy and momentum transfers.

1.2.2 Factorization

At least in principles, structure functions should be calculable. In practice, partons have a strong-coupling and many-body dynamics, which makes calculations very difficult.
However, a core property of QCD allows perturbation theory to be applied in some kinematical regions: the strong coupling constant decreases at high scales, and tends to zero as the scale tends to infinity. This phenomenon is referred to as *asymptotic freedom*.
As a consequence, perturbation theory can be applied to QCD at large scales.

Thanks to *factorization theorem*, the structure functions, which parametrize a physical cross section, are expressed as the convolution of perturbatively computable parton cross sections (coefficient functions), times parton distribution functions (pdfs); for example

$$F_2(x,Q^2) = x\cdot\Big(f_a\otimes\hat{\sigma}_a\Big) = x\cdot\int_x^1\frac{d\xi}{\xi}\sum_a f_a(\xi,Q^2)\cdot\hat{\sigma}_a\Big(\frac{x}{\xi},\alpha_S(Q^2)\Big) \tag{15}$$

where the index a is a parton label to be summed over all contributing quarks, antiquarks and gluons; $\otimes$ is a convolution. The factors have the following meaning:

- $f_a(\xi,Q^2)$ is a *parton distribution function* (pdf) which can be roughly understood as the probability of finding a parton inside the nucleon target with the momentum fraction x, at the scale Q^2;
- $\hat{\sigma}_a\big(\frac{x}{\xi},\alpha_S(Q^2)\big)$ is the *hard cross section*, i.e. the probability of the elastic scattering off the parton a.

To first order approximation,

$$F_2(x,Q^2) \approx x \sum_a e_a^2 \cdot f_a(x,Q^2) \quad , \tag{16}$$

the factorization formula is consistent with the parton model: the high energy scattering process is interpreted as essentially classical and incoherent. In eq.(15) this is found in the fact that the DIS cross section is computed by combining probabilities, rather than amplitudes.

The hard cross sections, involving only short-distance interactions, can be calculated by means of perturbative QCD, provided Q^2 is large enough, in the region of asymptotic freedom.
All long distances interactions of the DIS structure functions, that cannot be predicted by QCD, are factorized into the pdfs. Therefore, in order to be able to use the factorization equation to make prediction above high energetic processes we need to determine the parton distribution functions.

1.2.3 Parton Densities

The dependence of pdfs on the scale is weak: it is only logaritmic, and it can be understood as a consequence of the fact that quarks and gluons continuously interact with each other, emitting or splitting into new quark-antiquark pairs, and so the momentum distribution between the constituents of the nucleon is constantly changing. At high resolution these next to leading orders phenomena need to be taken into account: quarks and gluons turn out to be made of quarks and gluons themselves and nucleons are densely filled with partons.

The scale dependence of pdfs can be calculated in perturbation theory. If the shape of $f_a(\xi, Q^2)$ is known at a given scale Q_0^2, then it can be predicted with QCD for all other values of Q^2, by solving the *evolution equation* [11]:

$$Q^2 \frac{\partial f_a(x,Q^2)}{\partial Q^2} = \frac{\alpha_s}{2\pi} P_a^b \otimes f_b = \frac{\alpha_s}{2\pi} \int_x^1 \frac{\mathrm{d}\xi}{\xi} \sum_b P_a^b(\xi, \alpha_S) f_b(\frac{x}{\xi}, Q^2) \tag{17}$$

where $\{P_a^b\}$ are called splitting functions and can be perturbatively calculated order by order.

On the contrary, the x-dependence of pdfs cannot be predicted from first principles. They are determined by comparing experimental data (e.g. DIS cross sections) at a certain scale, to equation (15) or its analogue for other physical process. Several methods have been developed to this goal. In the next chapter we shall describe the approach proposed by the NNPDF collaboration.

2 Determination of Parton Distributions: the NNPDF Approach

During the last decades, the theory of strong interactions has evolved into precision physics, thanks to the wealth of experimental data coming especially from HERA and Tevatron.

A quantitative understanding of the phenomenology of quarks and gluons is made all the more necessary by the forthcoming experiments of proton-proton collisions at LHC, which aim at searches for new physics and require a solid understanding of the proton structure.

This involves both accurate perturbative computations at higher orders, and a precise determination of quantities, such as the parton distributions, which cannot be computed from first principles, together with an accurate estimation of their uncertainties.

Here we will concentrate on this second issue, considering the approach to pdf determination proposed by the NNPDF collaboration.
The method, based on a Monte Carlo approach, with neural networks used as universal unbiased interpolants, is designed to provide a faithful and statistically sound representation of the uncertainty on pdfs.

2.1 The Problem of Pdf Uncertainties

Until the turn of the century, standard sets of pdfs did not include uncertainties. However, pdfs were used to calculate observables that themselves had large theoretical uncertainties, and this shortcoming was not a problem. But with the need of quantitative tests of QCD, the issue of pdf uncertainties could not be sidestepped any longer: in fact, the uncertainty on the input pdfs is now found to be the leading term of the theoretical error on the predicted cross section [4].

As a first attempt to overcome the problem, the spread between different sets of pdfs was taken as an estimate of their uncertainty; an unreliable qualitative method, because the many possible sources of systematical bias are likely to be common to several parton determinations. It was abandoned when the accurate estimation of uncertainties became mandatory.

The first difficulty to be faced is that the problem of assessing pdfs is in fact mathematically ill-posed: we want to construct a probability measure on an infinite space (of functions), relying upon the knowledge of a finite set of experimental data.
The standard approach is that of projecting the infinite-dimensional problem onto a finite-dimensional space of parameters. A parametrization for the pdfs is assumed, based on some fixed functional form, and then parameters are tuned so that the computed observables fit the data. So the uncertainties are essentially error ellipsoids in the finite space of parameters.
However, the choice of a functional form is clearly a potential source of theoretical bias.

Another problem may be related to data incompatibilities. Indeed, many current parton fits effectively rescale all experimental uncertainties by some suitable large factor (*tolerance*); because this factor depends on the data set, a direct statistical interpretation of the results is lost. For example, benchmark studies [6] have shown that reducing the data set leads to results which are not compatible with those of a fit to the full data set.

In order to overcome these problems, a novel approach has been suggested by the NNPDF collaboration.

2.2 The NNPDF Approach

The general underlying strategy is twofold:

- a MC ensemble of replicas of the experimental data is generated, which can be interpreted as a sampling of the probability density on the space of physical observables, at the discrete points where data exist.
- then, neural networks are used to interpolate between these points: neural parton distributions are used to compute physical observables, which are then compared with the data to tune the best fit form of the pdfs.

As a consequence, the space of physical observables is mapped onto the space of pdfs: starting from the MC representation of the probability density at the points where data do exist, neural networks reproduce a representation of the probability density everywhere in the space of pdfs.
In other words, the pdfs with error that we obtain are themselves given as a MC

sample, so that any statistical property can be straightforwardly derived, e.g the average value of a function at some point is simply given by the average over the replicas at that point, the uncertainty is the variance, and so forth.

The viability of this procedure has been demonstrated initially by providing a determination of the proton and neutron structure functions [5]; then, in ref. [8] it has been used for the first time to calculate a quantity that cannot directly be measured, a non singlet parton distribution. The method has now been used to determine a full parton set.

2.2.1 Outline of the Strategy

The strategy adopted involves many steps, as pictured in fig. 1.

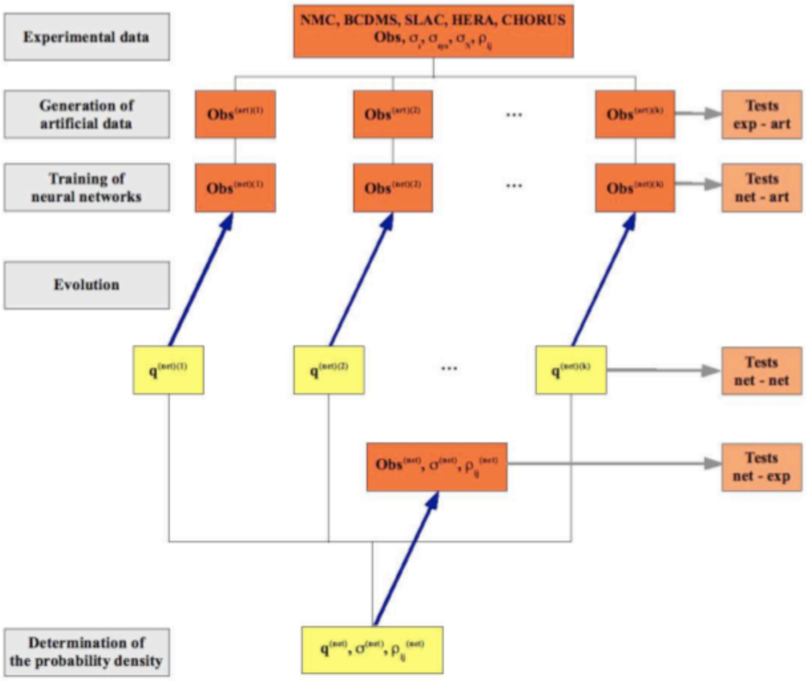
Experimental data
NMC, BCDMS, SLAC, HERA, CHORUS
Obs, σ_s, σ_{sys}, σ_N, ρ_{ij}
Generation of artificial data
$Obs^{(art)(1)}$
$Obs^{(art)(2)}$
...
$Obs^{(art)(k)}$
Tests exp - art
Training of neural networks
$Obs^{(net)(1)}$
$Obs^{(net)(2)}$
...
$Obs^{(net)(k)}$
Tests net - art
Evolution
$q^{(net)(1)}$
$q^{(net)(2)}$
...
$q^{(net)(k)}$
Tests net - net
$Obs^{(net)}$, $\sigma^{(net)}$, $\rho_{ij}^{(net)}$
Tests net - exp
Determination of the probability density
$q^{(net)}$, $\sigma^{(net)}$, $\rho_{ij}^{(net)}$

Figure 1: Schematic representation of the NNPDF method [12].

The first stage is the *generation of the pseudo-data* from a set of $N_{\rm dat}$ experimental data.

For each point, N_{rep} replicas are produced via MC method, using a Gaussian distribution described by the covariance matrix provided by the experimental collaborations. The number of replicas is chosen so that the sample reproduces the statistical features of the experiment, which can be checked using standard statistical methods ("Test exp-art" in the picture).

The second stage is the *construction of the parton sets*, and it involves several sub-steps.

First of all, a parton set contains a number N_{pdf} ($1 \leq N_{\mathrm{pdf}} \leq 13$) of pdfs, one for each contributing quarks and anti-quarks (12 at most) and one for the gluon distribution; in ref. [12] five pdfs have been used, for the *up* and *down* pairs of quarks and anti-quarks and for the gluon; in ref. [13], independent parametrization of the strange and anti-strange distributions has also been included.

Each pdf is parametrized using a neural network with a size and an architecture much larger than what would suffice: with a redundant parametrization, the fit is made independent of any assumption on the functional form.

The neural pdfs are expressed as functions of x, at a given scale Q_0. They are then evolved to the scale at which data are available, by means of perturbative QCD. Finally they are convoluted with the hard cross sections and used to reproduce the physical observables, to be compared with the experimental data.

The fitting procedure enters at this step: the set of computed observables is compared with each data replica, so that N_{rep} parton sets are eventually obtained. The fitting itself is performed with a *genetic algorithm*, a standard method in the context of neural networks. The figure of merit to be minimized is the χ^2 of the experimental points, computed by fully including the covariance matrix of the correlated experimental uncertainties.

Determining the optimal fit has posed a non trivial problem, because one would like to estimate the best fit without fitting statistical fluctuations of the data. The procedure implemented is the so-called *cross validation* method, well established in the context of neural networks: the idea is that the quality of the fit to data which have been used in the fitting procedure must be the same as the quality of the fit to data which have not been included in the fit.

At the end of the minimization, a MC sample of parton sets is obtained: for each pdf, $N_{\rm rep}$ replicas provide the corresponding probability density, so that uncertainties and averages can be assessed directly, by means of standard statistical tools ("test net-net"). The reliability of the results can be checked by the comparison of the final fit prediction with the original data ("test net-exp").

In the next section, we are going to delve deeper into some of the steps overviewed above. Specifically, we will concentrate on the Monte Carlo method and on the neural networks and their training.

2.2.2 Monte Carlo method

As mentioned above, a major problem for faithfully determining pdfs is that of computing the probability measure in the functional space of possible functions describing parton distributions.
In the NNPDF approach this measure is provided by a MC ensemble: first, an ensemble of artificial data is generated, in order to reproduce the statistical features of the experimental data. Then, a fit to each data replica is performed, so that a MC ensemble of pdfs is obtained, and the original probability density is mapped into the functional space of pdfs.

Let us describe in some detail the pseudo-data generation process. Artificial replicas of data points are distributed among a multi-Gaussian distribution centered on each data point; specifically, given a data point $F_p^{\rm exp}$, $N_{\rm rep}$ artificial points $F_p^{({\rm art})(k)}$ are generated as follows

$$F_p^{({\rm art})(k)} = S_{p,N} F_p^{\rm exp}\left(1 + \sum_{l=1}^{N_c} r_{p,l}^{(k)}\sigma_{p,l} + r_p^{(k)}\sigma_p^{\rm stat}\right) \tag{18}$$

and

$$S_{p,N}^{(k)} = \prod_{n=1}^{N_a}\left(1 + r_{p,n}^{(k)}\sigma_{p,n}\right)\prod_{n=1}^{N_r}\sqrt{1 + r_{p,n}^{(k)}\sigma_{p,n}} \quad . \tag{19}$$

where $\sigma_{p,l}$ are the correlated uncertainties, $\sigma_{p,n}$ are the normalization uncertainties and $\sigma_p^{\rm stat}$ is the statistical error of the experimental point. The variables $r_{p,l}^{(k)}$, $r_p^{(k)}$ and $r_{p,n}^{(k)}$ are univariate Gaussian random numbers that generate fluctuations of the pseudo-data around the central value of experimental data.
Whereas two points p and p' have correlated systematic uncertainties, the fluctuations that they generate are imposed to be the same, i.e. $r_{p,l}^{(k)} = r_{p',l}^{(k)}$.

2.2.3 Neural Networks

While current parton sets are obtained using a set of *a priori* selected functions, parametrized by a small number of physically motivated parameters, the NNPDF approach to parametrization of parton densities is novel, based on the use of an unbiased basis of functions, described by a large number of parameters. On the one hand, this choice guarantees the lack of theoretical bias that plagues other methods; on the other, it poses practical problems in the fitting procedure: first of all the minimum has to be found in a very large space of parameters, and then a redundant parametrization runs the risk of accommodating not only the smooth shape of the true pdfs but also the random fluctuations of the data about them. The first problem is solved by using a genetic algorithm for minimization, the second problem is solved by the choice of a stopping criterion based on cross-validation.

Neural networks parametrization Artificial neural networks provide a particularly convenient set of interpolating functions, able to well approximate incomplete and noisy data; in the limit of infinite size they can reproduce any continuous function.

A neural network is a set of interconnected units, the *neurons*. Each neuron ξ_i is a real number, determined as a function (*activation*) of the neurons connected to it:

$$\xi_i = g\left(\sum_j \omega_{ij}\xi_j - \theta_i\right) \tag{20}$$

where $\omega_{ij} \in \mathbb{R}$ is the *weight* describing the connection (*synapsis*) between a pair of neurons, and θ_i is referred to as *threshold*.

The activation function g is in general non-linear, in order to enable neural networks to reproduce nontrivial functions; a common choice for the activation function is the sigmoid:

$$g(x) \equiv \frac{1}{1 + e^{-\beta x}} \tag{21}$$

Here we consider in particular the *multilayer feed-forward neural networks* (depicted in fig. 2): these are organized in ordered layers whose neurons only receive

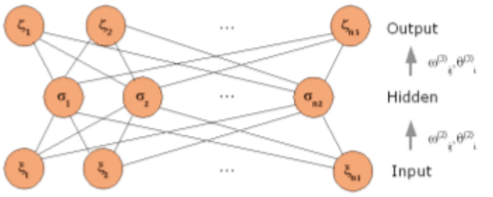

Figure 2: schematic picture of multilayer feed-forward neural networks [8].

input from the previous layer. The following recursive relation holds:

$$\xi_i^l = g\left(\sum_{j=1}^{n_l-1} \omega_{ij}^{(l-1)} \xi_j^{l-1} - \theta_i^{(l)}\right) \qquad i = 1, ..., n_l; \quad l = 2, ..., L \tag{22}$$

for L layers with $n_1, ..., n_L$ neurons respectively.
We can say that a multilayer neural network is a non-linear map $\mathbb{R}^{n_1} \longrightarrow \mathbb{R}^{n_L}$, parametrized by weights, thresholds and the activation function.

In the limit of infinite size, a neural network can be trained to reproduce any function, by tuning its parameters. In practice, the reason why neural networks can be considered as unbiased universal approximants lies in the fact that they can be built to be redundant: one may check that, for a given problem, adding or removing a neuron has little or no effect on the final output. The architecture used by NNPDF is 2-5-3-1: it turns out that using x and $\ln x$ as simultaneous inputs improves the efficiency of the minimization, with respect to the option of having more neurons in the hidden layers.

Genetic algorithm minimization The state of each neural network is determined by the values of its weights ω_{ij} and thresholds θ_i, which can be arranged in a vector of N_{par} parameters $\omega = (\omega_1, ..., \omega_{N_{\text{par}}})$ (*weight vector*); the task of fitting the data consists of simultaneously tuning the weight vectors of N_{pdf} neural networks, by minimizing a suitable figure of merit. The following error function is used:

$$E^{(k)}[\omega] = \frac{1}{N_{\mathrm{dat}}} \sum_{i,j=1}^{N_{\mathrm{dat}}} \left(F_i^{(\mathrm{art})(k)} - F_i^{(\mathrm{net})(k)}\right)\left((\overline{\mathrm{cov}}^{(k)})^{-1}\right)_{ij}\left(F_j^{(\mathrm{art})(k)} - F_j^{(\mathrm{net})(k)}\right) \qquad (23)$$

where $F_i^{(\mathrm{net})}$ is the physical observable, computed from neural pdfs, corresponding to the i-th data point.

The minimization of the error function $E^{(k)}[\omega]$ leads to one replica of the parton set, whereas the quality of the global fit can be estimated by the χ^2 computed from the average over the whole final sample of parton sets:

$$\chi^2 = \frac{1}{N_{\mathrm{dat}}} \sum_{i,j=1}^{N_{\mathrm{dat}}} \left(F_i^{(\mathrm{exp})} - \langle F_i^{(\mathrm{net})}\rangle_{\mathrm{rep}}\right)\left((\mathrm{cov})^{-1}\right)_{ij}\left(F_j^{(\mathrm{exp})} - \langle F_j^{(\mathrm{net})}\rangle_{\mathrm{rep}}\right) \quad . \qquad (24)$$

Regarding the minimization technique, a genetic algorithm is used, which turns out to be convenient when seeking a minimum in a wide space with potentially several local minima: in fact, it explores simultaneously many regions of the parameters space, handling a population of solutions rather than traversing a path in the space of solutions, thereby eluding the danger of being trapped in local minima.

For each replica, N_{mut} copies of the N_{pdf} weight vectors are generated, forming an initial set; the following steps are then performed repeatedly, each new cycle referred to as *generation*:

1. the initial set is replicated into N_{cop} copies;
2. (*mutation*) for each copy, $N_{\mathrm{mut}}\times N_{\mathrm{pdf}}$ mutations are performed: one randomly chosen element of each weight vector is replaced by the new value

   $$\omega_n^{(i,j)} = \omega_n^{(i,j)} + \eta_{ij}\left(r - \frac{1}{2}\right) \qquad i = 1, ..., N_{\mathrm{pdf}}; \quad j = 1, ..., N_{\mathrm{mut}} \qquad (25)$$

   where r is a random number between 0 and 1, and η_{ij} are the reaction rates, free parameters of the minimization, tuned to optimize the efficiency of the procedure; namely, they are adjusted dynamically along the fit, decreasing as the minimum is being approached;
3. (*selection*) the copy whith the lowest value of the error function is selected and used to replace the initial set of weight vectors.

The whole process is iterated until the weight vectors which yields to the lowest value of the error function (in each generation) meet a suitable criterion of convergence, that we are now going to discuss.

Cross validation method As we have briefly mentioned before, because of the flexibility of the neural network parametrization it is possible for it to fit random statistical fluctuations of the data (overlearning). Suppose, for example, to have data for the same quantity at two different but very close values of x: a fit that goes through the central values of both measurements is possible, but, assuming that the data are taken at infinitesimally close x, this would potentially lead to a discontinuous behavior of the observable, which is certainly unphysical. This would be found if one stopped at the absolute minimum of the figure of merit.

The best fit should be the weighted average of the infinitesimally close measurements. A way of determining the best fit which is free of this problem is based on the cross validation method.
The procedure consists in randomly dividing each set of replicas into two subsets of $N_{\text{dat}}\,/2$ data, a *fitting set* and a *validating set*; while training the neural networks, two error functions are calculated by comparing the theoretical observables separately with each one of the subsets.
The fit is stopped when the error function of the validation set stops decreasing; if the χ^2 of the fitting set can be further minimized it means that the overlearning regime has been entered.

3 The Treatment of Correlated Uncertainties in s global pdf analysis

The systematic errors arising within each data set are often highly correlated. Some groups of pdf fitting, such as NNPDF, CTEQ, Alekhin, treat correlated uncertainties properly, by fully including the covariance matrixes of the experiments; other groups do not (MSTW).

In the NNPDF approach, correlations appear twice: first, at the level of Monte Carlo generation of data replicas, then at the level of the χ^2-minimization in the fitting procedure.

In ref. [10], it is argued that results on global pdf analysis are essentially unchanged whether one treats the systematic correlated uncertainties properly or if one just neglects correlations between experimental data points.

In practice, this means that instead of minimizing the error function defined in eq. (23),

$$E^{(k)}_{\rm cme}[\omega] = \frac{1}{N_{\rm dat}} \sum_{i,j=1}^{N_{\rm dat}} \left(F_i^{({\rm art})(k)} - F_i^{({\rm net})(k)}\right)\left((\overline{\rm cov}^{(k)})^{-1}\right)_{ij}\left(F_j^{({\rm art})(k)} - F_j^{({\rm net})(k)}\right)$$

one is minimizing a diagonal error function, in which correlations are neglected and uncertainties are simply added in quadrature:

$$E^{(k)}_{\rm diag}[\omega] = \frac{1}{N_{\rm dat}} \sum_{i=1}^{N_{\rm dat}} \frac{\left(F_i^{({\rm exp})} - F_i^{({\rm net})}\right)^2}{\sigma_i^{({\rm stat}),2} + \sigma_i^{({\rm sys}),2}} \quad . \qquad (26)$$

In order to clarify this statistical issue, we have varied some of the default settings of the NNPDF parton analysis, and studied their impact on the pdfs.

After a brief description of the experimental data used in this analysis, we shall present our results.

The Data Set We refer to the data set used in the NNPDF global analysis presented in ref. [14] (NNPDF1.2). It is a comprehensive set of experimental data coming from DIS with various lepton beams and nucleon targets. The kinematical coverage and the main statistical features of data are summarized in table. 1 and fig. 3.

Experiments	Set	Points	σ_{stat} (%)	σ_{sys} (%)	σ_{tot} (%)
NMC-pd		260	2	0.4	2.1
NMC		288	3.7	2.3	5
SLAC					
	SLACp	211	2.7	0	3.6
	SLACd	211	2.5	0	3.2
BCDMS					
	BCDMSp	351	3.2	2	5.5
	BCDMSd	254	4.5	2.3	6.6
ZEUS					
	Z97lowQ2	80	2.8	3	4.9
	Z97NC	160	6.2	3.1	7.7
	Z97CC	29	33.6	5.5	34.2
	Z02NC	92	12.7	2.3	13.2
	Z02CC	26	39.6	6.3	40.2
	Z03NC	90	7.7	3.3	9.1
	Z03CC	30	29.9	6.7	31
H1					
	H197mb	67	3.8	2.1	4.9
	H197lowQ2	80	2.7	2.5	4.2
	H197NC	130	12.5	3.2	13.3
	H197CC	25	29.3	4.5	29.8
	H199NC	126	14.9	2.8	15.5
	H199CC	28	27.1	3.8	27.6
	H199NChy	13	8.7	1.9	9.2
	H100NC	147	9.4	3.2	10.4
	H100CC	28	21.3	3.8	21.8
CHORUS					
	CHORUSnu	607	4.2	6.4	11.2
	CHORUSnb	607	13.8	7.8	18.7
FLH108		8	48.8	48.6	69.2
NTVDMN					
	NTVnuDMN	45	17.2	1	17.7
	NTVnbDMN	45	26.5	0	26.5
ZEUS-H2					
	Z06NC	90	4.1	3.7	6.6
	Z06CC	37	25.3	14.1	31.6

Table 1: main features of the data set. Data are divided into experiments and sets; different sets within the same experiment present correlated systematics. The table shows the number of data points and the averaged uncertainties, pointing out the contribution of the statistical and systematic uncertainties.

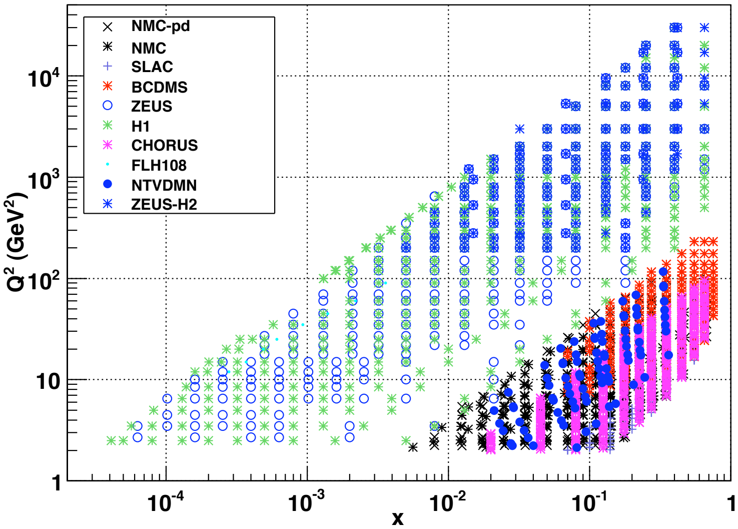

Figure 3: kinematical coverage of the data set employed in this work ([14]); Q^2 is plotted versus x on a logaritmic scale.

The covariance matrix for each experiment can be computed from knowledge of statistical, systematic and normalization uncertainties:

$$\mathrm{cov}_{ij} = \left(\sum_{l=1}^{N_C} \sigma_{i,l}\sigma_{j,l} + \sum_{n=1}^{N_a} \sigma_{i,n}\sigma_{j,n} + \sum_{m=1}^{N_r} \sigma_{i,m}\sigma_{j,m} + \delta_{ij}\sigma_{i,S}^2 \right) F_i F_j \tag{27}$$

where i and j label the experimental points, $\sigma_{i,l}$ are the correlated uncertainties, $\sigma_{i,n}$ and $\sigma_{i,m}$ are respectively the absolute and relative normalization uncertainties and $\sigma_{i,S}$ are the statistical uncertainties.
The total uncertainty for the i-th point is defined as

$$\sigma_{i,(\mathrm{tot})} = \sqrt{\sigma_{i,S}^2 + \sigma_{i,C}^2 + \sigma_{i,N}^2} \tag{28}$$

with $\sigma_{i,C}$ and $\sigma_{i,N}$ are the sum of all correlated and all normalization uncertainties respectively.

3.1 Procedure and Results

Our analysis is based upon samples of 100 replicas, which is a sufficient number in order to reproduce central values and uncertainties to an accuracy of about 1% ([12]).

We have produced two different parton sets, one that reproduces NNPDF global analysis, the other which is close to the treatment of systematic uncertainties of the MSTW group, namely neglecting correlations and adding all errors in quadrature. In fact, MSTW do include correlated systematics for HERA experiments and some other, but neglects them for all fixed target experiments.

The first, referred to as *CME set*, was obtained using standard NNPDF settings, so that correlations are fully retained. The fit is compatible with that of ref. [13], i.e. distances between the two sets are of order 1 (the distance is defined in eq. (31) in sec. 3.1.2).

On the contrary, correlations have been completely neglected in producing the second set, referred to as *diagonal set*, both at the level of MC generation and in the minimization.
The Monte Carlo sample was produced using, instead of eq. (18), the simplified formula:

$$F_p^{(\mathrm{art})(k)} = S_{p,N} F_p^{\mathrm{exp}} \Big(1 + r_p^{(k)} \sigma_{p,\mathrm{tot}} \Big) \quad , \tag{29}$$

for the k-th replica of point F_p^{exp}; $\sigma_{p,\mathrm{tot}}$ is defined in (28). Turning off correlations at this level means that we allow data to fluctuate within the same interval as before, in terms of absolute values, but each point independently with all the others. Similarly, at the level of χ^2-minimization, we have used eq. (26), where systematics are simply added in quadrature.

Let us note that if we only changed the settings of the MC generation, retaining correlations during the minimization, we would expect the central values of the pdfs to be the same of the cme pdfs, but with different uncertainties. On the contrary, if we performed a diagonal-χ^2-minimization over the same set of replicas that we use to produce the cme parton set, we would then expect different central values, but the same uncertainties.
Unfortunately this second analysis cannot be successfully carried out, because, as we will show, the difference between the uncertainties of the pdfs are too small to be distinguished from the statistical fluctuations of data.

3.1.1 Comparison of the Fits

The comparison of the fits has been made using two statistical estimators, the $\chi^2_{\rm cme}$given in eq. (24), and the $\chi^2_{\rm diag}$, defined as

$$\chi^2_{\rm diag} = \frac{1}{N_{\rm dat}} \sum_{i=1}^{N_{\rm dat}} \frac{\left(F_i^{\rm (exp)} - \langle F_i^{\rm (net)} \rangle_{\rm rep}\right)^2}{\sigma_i^{\rm (stat),2} + \sigma_i^{\rm (sys),2}} \quad . \tag{30}$$

The values of these χ^2s, for each experiment and each set, are shown in table 2, together with a comparison, whenever possible, with the MSTW08 fit ([10]).

Let us consider the two fits that we have performed.
First of all, comparing the χ^2s for the full set of data, we notice that the quality of the two fits is essentially unchanged if one considers the $\chi^2_{\rm cme}$; on the contrary the values of the $\chi^2_{\rm diag}$ are significantly different. This suggests that minimizing a diagonal χ^2 artificially gives larger weight to experiments with smaller systematics: indeed, if large systematics are added in quadrature, they causes the χ^2 to decrease. However, this does not lead to differences of the $\chi^2_{\rm cme}$, because the fit to experiments with larger systematics deteriorates.

In fact, we see that for experiments with significant systematics (NMC, BCDMS, CHORUS) we can make the same comment: the $\chi^2_{\rm cme}$ is basically unchanged, while the $\chi^2_{\rm diag}$ is quite different for the two fits.
In general, we see that for these data sets the value of the $\chi^2_{\rm diag}$ is rather smaller than that of $\chi^2_{\rm cme}$, within the same fit; this is because adding large systematics in quadrature cuases the χ^2 to decrease. This phenomenon is very significant for FLH108 and Z06CC which have the largest systematics.

For some of the experiments with negligible systematics (NMC-pd, NTVDMN) we see that the χ^2s within the same fit are comparable, but the values are very different if one looks at the two fits. On the one hand, correlations are small and their inclusion doesn't change the quality of the fit when measured with the two different estimators. On the other hand, the qualities of the two fits are different because the global fit has been significantly modified: we notice that NMC-pd and NTVDMN explore the same kinematical region as NMC, BCDMS and CHORUS, for which, we have seen, the inclusion of systematics does make a difference; neglecting correlations for these experiments causes the global fit to change so

NNPDF method						MSTW08	
		CME fit		Diagonal fit			
Experiment	Set	χ^2_{diag}	χ^2_{cme}	χ^2_{diag}	χ^2_{cme}	χ^2_{diag}	(corresponding label)
TOT (all exp)		0.988	1.323	0.844	1.321	0.942	TOT (all exp)
NMC-pd		1.965	1.457	1.167	1.155	0.878	NMC $\mu n/\mu p$
NMC		1.006	1.659	1.078	1.76	0.984	NMC $\mu p F_2$
SLAC		0.836	1.185	1.008	1.406		
	SLACp	1.018	1.307	1.132	1.525	0.811	SLAC ep F_2
	SLACd	0.651	0.912	0.882	1.275	0.684	SLAC ed F_2
BCDMS		0.777	1.646	0.552	1.604		
	BCDMSp	0.873	1.808	0.617	1.703	1.117	BCDMS $\mu p F_2$
	BCDMSd	0.648	1.296	0.465	1.23	1.258	BCDMS $\mu d F_2$
ZEUS		0.770	1.055	0.742	1.048		
	Z97lowQ2	0.474	1.294	0.434	1.367	0.597	ZEUS 96-97 e^+pNC
	Z97NC	0.718	1.125	0.669	1.106		
	Z97CC	0.912	0.800	1.021	0.894		
	Z02NC	0.798	0.767	0.763	0.733		
	Z02CC	0.619	0.592	0.593	0.569		
	Z03NC	0.975	1.104	0.907	1.012		
	Z03CC	1.131	1.001	1.259	1.115		
H1		1.020	1.053	0.997	1.028		
	H197mb	0.861	1.298	0.877	1.33	0.656	H1MB97e^+pNC
	H197lwQ2	0.666	0.948	0.774	0.97	0.500	H197lowQ^2 96-97 e^+pNC
	H197NC	1.071	0.903	0.986	0.852		
	H197CC	0.758	0.764	0.831	0.824		
	H199NC	1.229	1.109	1.171	1.068	0.968	H1 highQ^2 98-99 e^-pNC
	H199CC	0.621	0.646	0.644	0.668		
	H199NChy	0.333	0.361	0.326	0.353		
	H100NC	1.208	1.172	1.120	1.102	0.891	H1 highQ^2 99-00 e^+pNC
	H100CC	1.122	1.013	1.311	1.146	1.036	H1 99-00 e^+pCC
CHORUS		1.018	1.380	0.745	1.392		
	CHORUSnu	1.082	1.449	0.628	1.403		
	CHORUSnb	0.954	1.178	0.861	1.254		
FLH108		0.984	1.729	0.946	1.7		
NTVDMN		0.869	0.692	1.094	0.984		
	NTVnuDMN	1.061	0.763	0.445	0.421		
	NTVnbDMN	0.667	0.660	1.774	1.618		
ZEUS-H2		1.392	1.509	1.373	1.512		
	Z06NC	1.691	1.495	1.667	1.472		
	Z06CC	0.664	1.230	0.659	1.252		

Table 2: comparison of the quality of the NNPDF fits and the MSTW08 fit. The values of the χ^2s for MSTW08 com from ref. [10].

much that NMC-pd and NTVDMN, despite not being affected by systematics themselves, are not fitted the same way.

HERA and ZEUS, which are collider experiments (the others are fixed-target experiments), have systematics that are negligeble if compared to the statistical uncertainty. In this case, we see that the difference between the χ^2s within the same fit are quite similar, and so are the qualities of the two fits. In this case, correlations have really a negligible effect, because HERA and ZEUS cover a kinematical region that is not controlled by any other data.
An exception is in the low Q^2 sets of both ZEUS and H1 (Z97lowQ^2, H197mb, H107lowQ^2), for the values of the $\chi^2_{\rm diag}$ and of the $\chi^2_{\rm cme}$ are very different and the same comment of the global set can be made.

Similarly, let us compare the global quality of the NNPDF fits and of the MSTW08 fit. For the three fits, we have calculated the total $\chi^2_{\rm diag}$ restricted to the data sets that we compare:

	CME fit	diagonal fit	MSTW08 fit
$\chi^2_{\rm diag}$	0.994	0.851	0.903

We see that the NNPDF fit quality is comparable to MSTW08 if we compute the $\chi^2_{\rm diag}$, and actually rather better if the $\chi^2_{\rm diag}$ is minimized instead of the default $\chi^2_{\rm cme}$ used in all published NNPDF fit. This is to be expected given that NNPDF data set is a subset of purely DIS data, while MSTW analysis is made on a wider data set.

3.1.2 Comparison between the CME and the diagonal Parton Sets

Up to now we have discussed the fits, and we have seen that, on average, the quality of the two fits compared is similar. However, note that, even when the χ^2 is similar, this doesn't imply that the fitted pdfs are the same.

In order to quantify this statement, we have calculated the distances between the pdfs coming from the CME fit and the diagonal fit.

The statistical estimator that we have used for this analysis, the *distance*, is defined as the difference in values between two different predictions for some quantity q obtained from two different ensembles of pdfs, measured in units of the sum in quadrature of their uncertainties [12]:

$$\langle d[q] \rangle = \sqrt{\left\langle \frac{\left(\langle q_i \rangle_{(1)} - \langle q_i \rangle_{(2)} \right)^2}{\sigma^2[q_i^{(1)}] + \sigma^2[q_i^{(2)}]} \right\rangle_{\mathrm{dat}}} \quad , \tag{31}$$

where

$$\langle q_i \rangle_{(n)} = \frac{1}{N_{\mathrm{rep}}^{(n)}} \sum_{k=1}^{N_{\mathrm{rep}}^{(n)}} q_{ik}^{(n)} \tag{32}$$

and

$$\sigma^2[q_i^{(n)}] = \frac{1}{N_{\mathrm{rep}}^{(n)}(N_{\mathrm{rep}}^{(n)} - 1)} \sum_{k=1}^{N_{\mathrm{rep}}^{(n)}} \left(q_{ik}^{(n)} - \langle q_i \rangle_{(n)} \right) \tag{33}$$

is the standard deviation. The analogue estimator for the uncertainty, $\langle d[\sigma] \rangle$, is obtained from eq. (31) substituting to q_i the uncertainty

$$\sigma_i^{(\mathrm{net})} = \sqrt{\frac{N_n}{N_{(n)} - 1} \left(\langle q_i^2 \rangle_{(n)} - \langle q_i \rangle_{(n)}^2 \right)} \quad . \tag{34}$$

If two determinations of q have a distance d, it means that their difference is equal to the sum in quadrature of their respective uncertainties.
It should be noticed that averages over N_{rep} replicas have a standard deviation that is smaller, by a factor $\sqrt{N_{\mathrm{rep}}}$, than the standard deviation of each replica. The distance between two sets of N_{rep} pdfs refers to an average, which means that the central values of these pdfs differ by $1/\sqrt{N_{\mathrm{rep}}}$ of the sum in quadrature of their standard deviations.

The distances between pdfs of the CME set and of the diagonal set are shown in table 3. They have been calculated twice, in the kinematical region where data are abundant and pdfs are mostly controlled by data (“Data”), and in the kinematical region where pdfs have been mostly extrapolated (“Extrapolation”).

These results are also shown in the plots of the pdfs that we provide below: the CME pdfs, obtained retaining correlations, are superimposed to those obtained adding uncertainties in quadrature, the diagonal pdfs.

The pdfs compared are the seven independent pdfs included in the NNPDF global analysis of ref. [14]. They are given by the following linear combinations:

	Data	Extrapolation
$\Sigma(x, Q_0^2)$	$0.5 \cdot 10^{-3} < x < 0.1$	$0.1 \cdot 10^{-4} < x < 0.1 \cdot 10^{-03}$
$\langle d[q] \rangle$	4.62 ± 0.07	1.64 ± 0.08
$\langle d[\sigma] \rangle$	1.44 ± 0.03	1.53 ± 0.08
$g(x, Q_0^2)$	$0.5 \cdot 10^{-3} < x < 0.1$	$0.1 \cdot 10^{-4} < x < 0.1 \cdot 10^{-3}$
$\langle d[q] \rangle$	1.06 ± 0.05	0.97 ± 0.05
$\langle d[\sigma] \rangle$	1.71 ± 0.04	1.20 ± 0.06
$T_3(x, Q_0^2)$	$0.5 \cdot 10^{-1} < x < 0.75$	$0.1 \cdot 10^{-2} < x < 0.1 \cdot 10^{-1}$
$\langle d[q] \rangle$	2.66 ± 0.07	1.17 ± 0.07
$\langle d[\sigma] \rangle$	1.28 ± 0.05	0.77 ± 0.06
$V(x, Q_0^2)$	$0.1 < x < 0.6$	$0.3 \cdot 10^{-2} < x < 0.3 \cdot 10^{-1}$
$\langle d[q] \rangle$	1.53 ± 0.09	1.62 ± 0.10
$\langle d[\sigma] \rangle$	1.07 ± 0.05	0.92 ± 0.06
$\Delta_S(x, Q_0^2)$	$0.1 < x < 0.6$	$0.3 \cdot 10^{-2} < x < 0.3 \cdot 10^{-1}$
$\langle d[q] \rangle$	2.91 ± 0.11	1.65 ± 0.10
$\langle d[\sigma] \rangle$	1.25 ± 0.06	1.31 ± 0.04
$S_P(x, Q_0^2)$	$0.5 \cdot 10^{-3} < x < 0.1$	$0.1 \cdot 10^{-4} < x < 0.1 \cdot 10^{-3}$
$\langle d[q] \rangle$	4.19 ± 0.08	2.08 ± 0.09
$\langle d[\sigma] \rangle$	1.58 ± 0.04	1.53 ± 0.10
$S_M(x, Q_0^2)$	$0.1 < x < 0.6$	$0.3 \cdot 10^{-2} < x < 0.3 \cdot 10^{-1}$
$\langle d[q] \rangle$	1.37 ± 0.05	1.79 ± 0.09
$\langle d[\sigma] \rangle$	1.59 ± 0.06	2.28 ± 0.08

Table 3: distances between *diagonal set* and *CME set*, at the reference scale $Q_0^2 = 2\text{GeV}^2$, in two different kinematical regions.

1. the singlet distribution, $\Sigma(x) \equiv \Sigma_{i=1}^{n_f}(q_i(x) + \bar{q}_i(x))$,
2. the gluon, $g(x)$,
3. the total valence, $V(x) \equiv \Sigma_{i=1}^{n_f}(q_i(x) - \bar{q}_i(x))$,
4. the nonsinglet triplet, $T_3(x) \equiv (u(x) + \bar{u}(x)) - (d(x) - \bar{d}(x))$,
5. the sea asymmetry distribution, $\Delta_S(x) \equiv \bar{d}(x) - \bar{u}(x)$,
6. the strange plus, $S_P(x) \equiv s(x) + \bar{s}(x)$,
7. the strange minus, $S_M(x) \equiv s(x) - \bar{s}(x)$,

where $n_f = 6$ is the number of flavors. The reference scale employed is $Q_0^2 = 2\text{GeV}^2$.

On average, the two parton sets are comparable within one σ. While distances between central values in the data region range between $\langle d[q] \rangle = 1.06$ for the gluon distribution, and $\langle d[q] \rangle = 4.62$ for the singlet distribution, we see that uncertainties and central values in the extrapolation region generally show a good agreement. Indeed, in the extrapolation region the size of uncertainties mostly reflects the lack of information and it is thus only loosely affected by the treatment of uncertainties in the data region.

With the graphs, it is made more apparent that even if the general trend is that pdfs do not differ for a remarkable amount, the averaged distances shown in table 3 may hide local behaviors where pdfs in fact differ more significantly.

The largest differences are observed for pdfs in the valence region, consistent with the observation that the treatment of systematics mostly affects the impact of BCDMS, NMC, CHORUS and NTVDMN.
This is particularly pronouced for the singlet and strange plus distributions, for which the diagonal pdf and the CME pdf differ for about half σ: central values differ for almost one σ in the region between $x = 0.1$ and $x = 0.3$.

The nonsinglet triplet and the sea asymmetry distributions show a more stable behavior.
For the nonsinglet triplet, we also notice that the diagonal pdf is underestimated, while the general trend is that of diagonal pdfs being overestimated. Is it possible that the effects of unretained systematics cancel out?

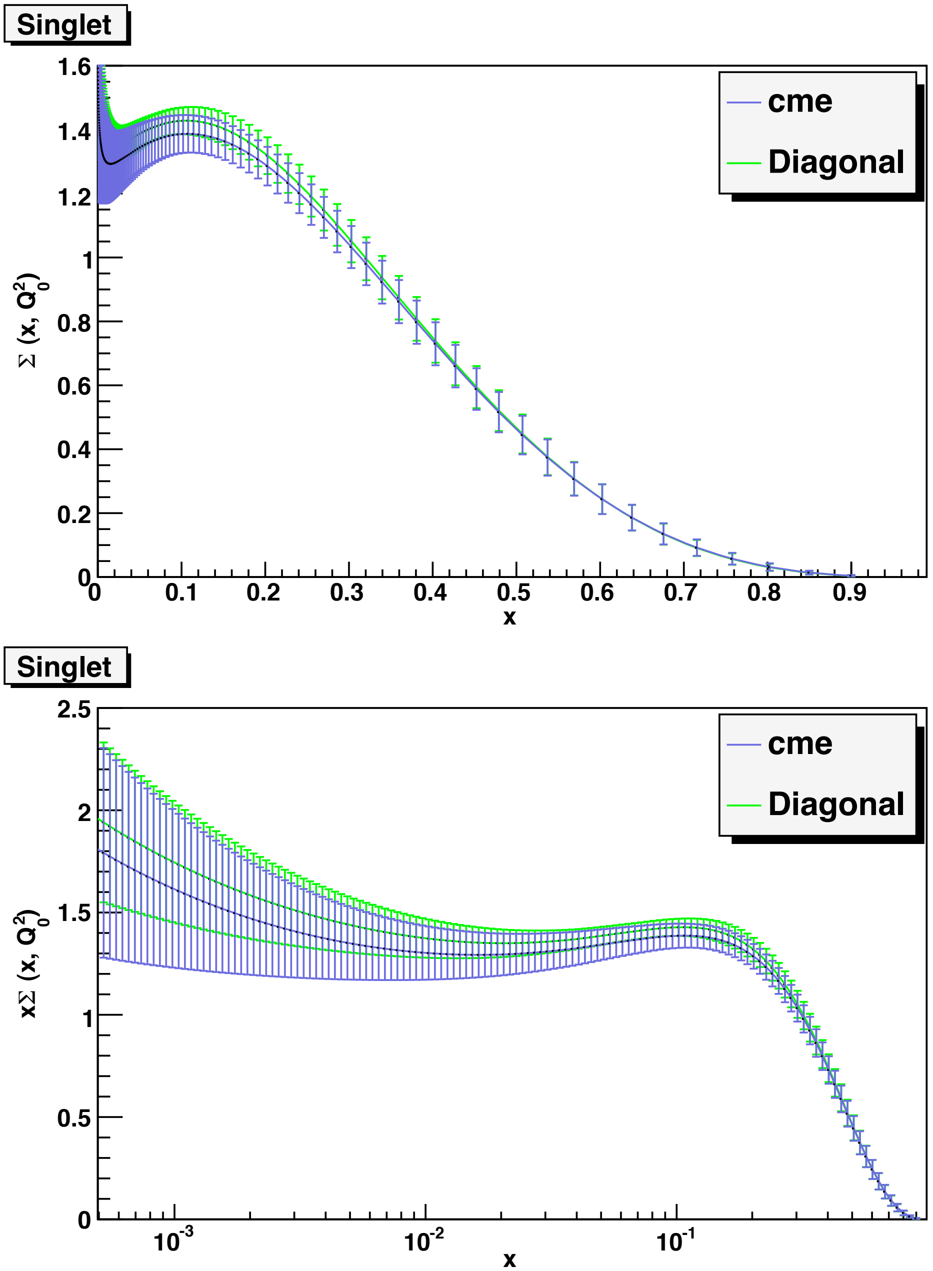

Figure 4: the singlet distribution at the reference scale $Q_0^2 = 2\text{GeV}^2$, plotted versus x on a linear (left) or a logaritmic scale (right).

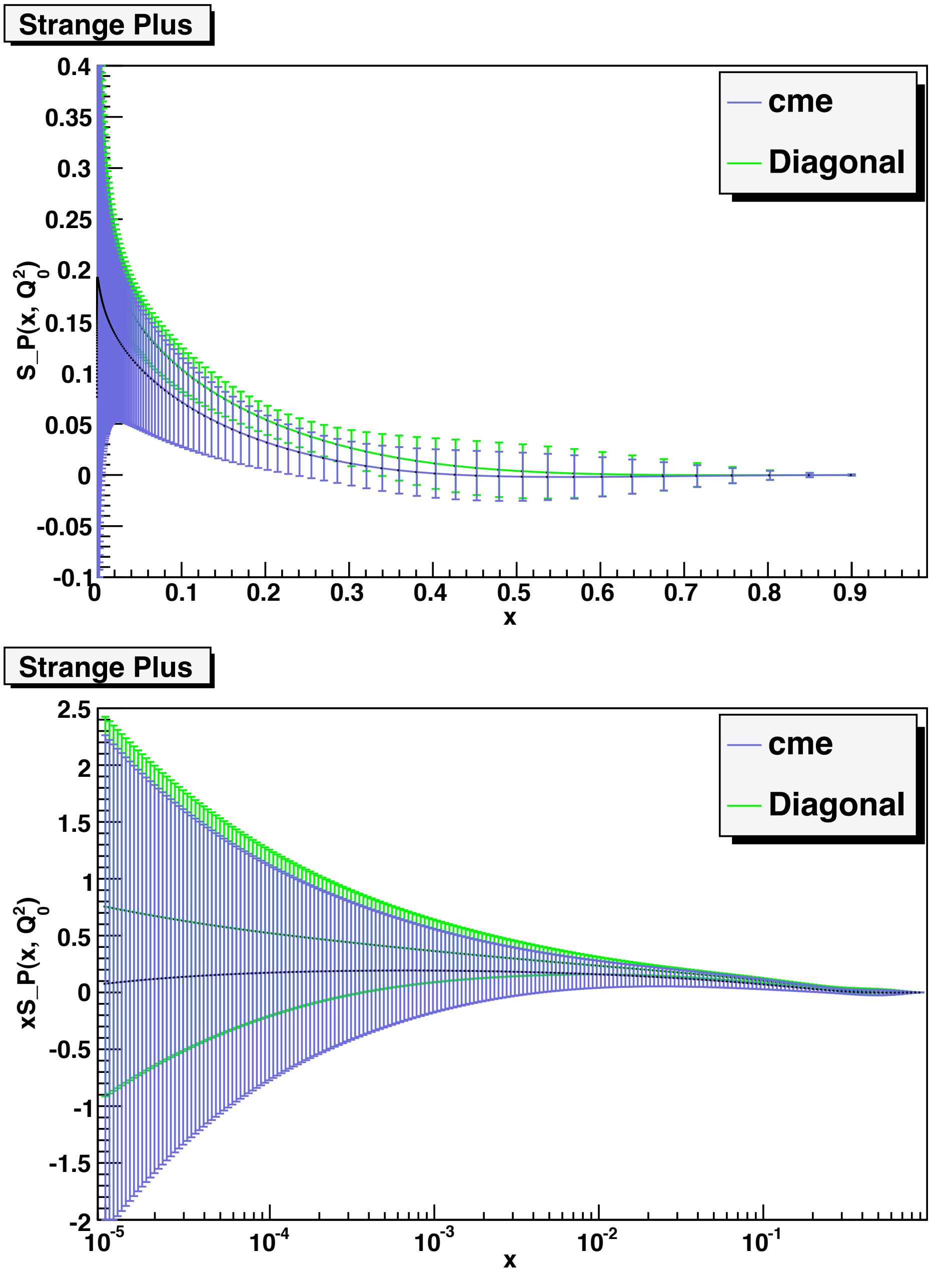

Figure 5: the strange plus at the reference scale $Q_0^2 = 2\text{GeV}^2$, plotted versus x on a linear (left) or a logaritmic scale (right).

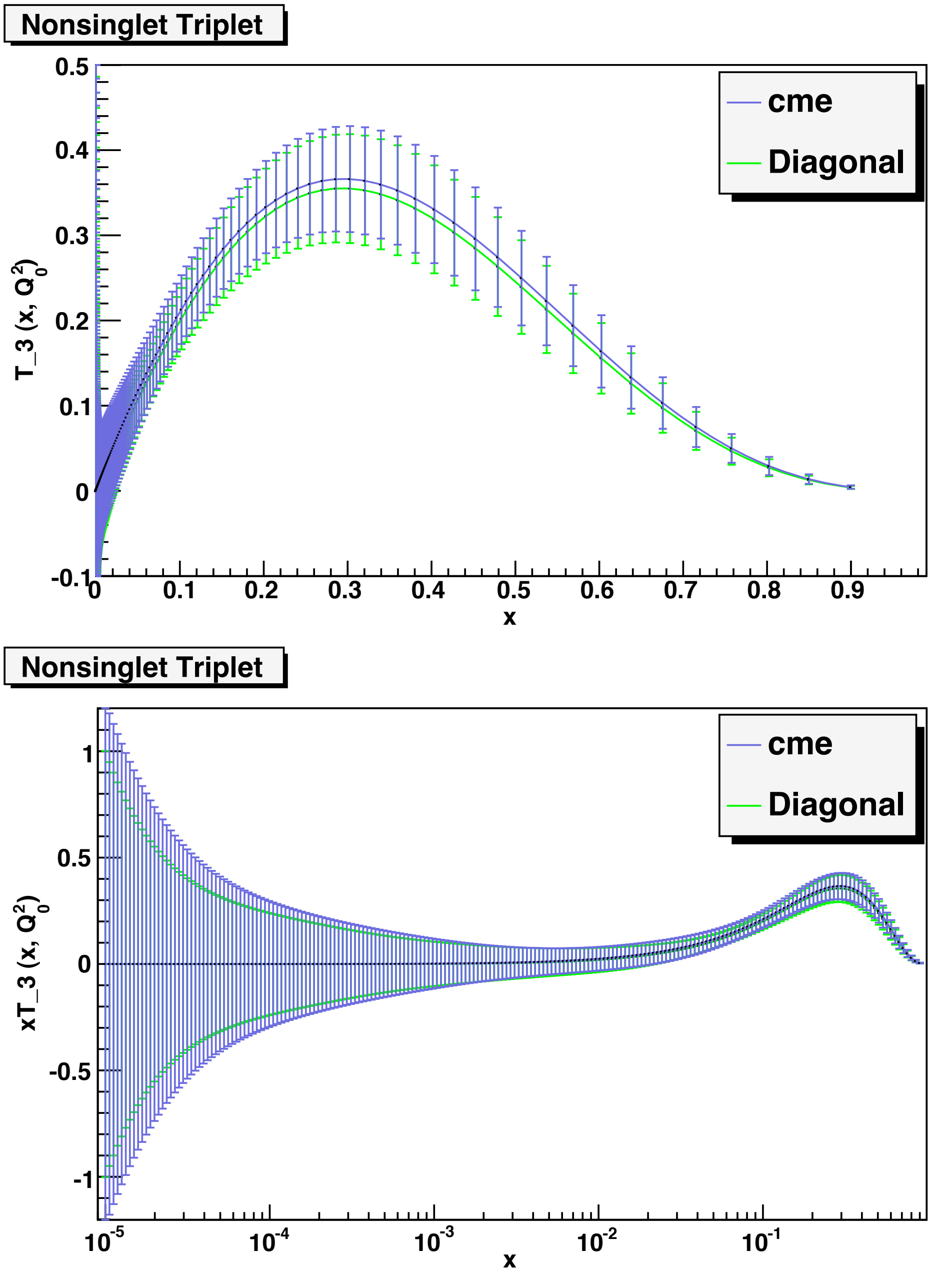

Figure 6: the nonsinglet triplet at the reference scale Q_0^2 = 2GeV^2, plotted versus x on a linear (left) or a logaritmic scale (right).

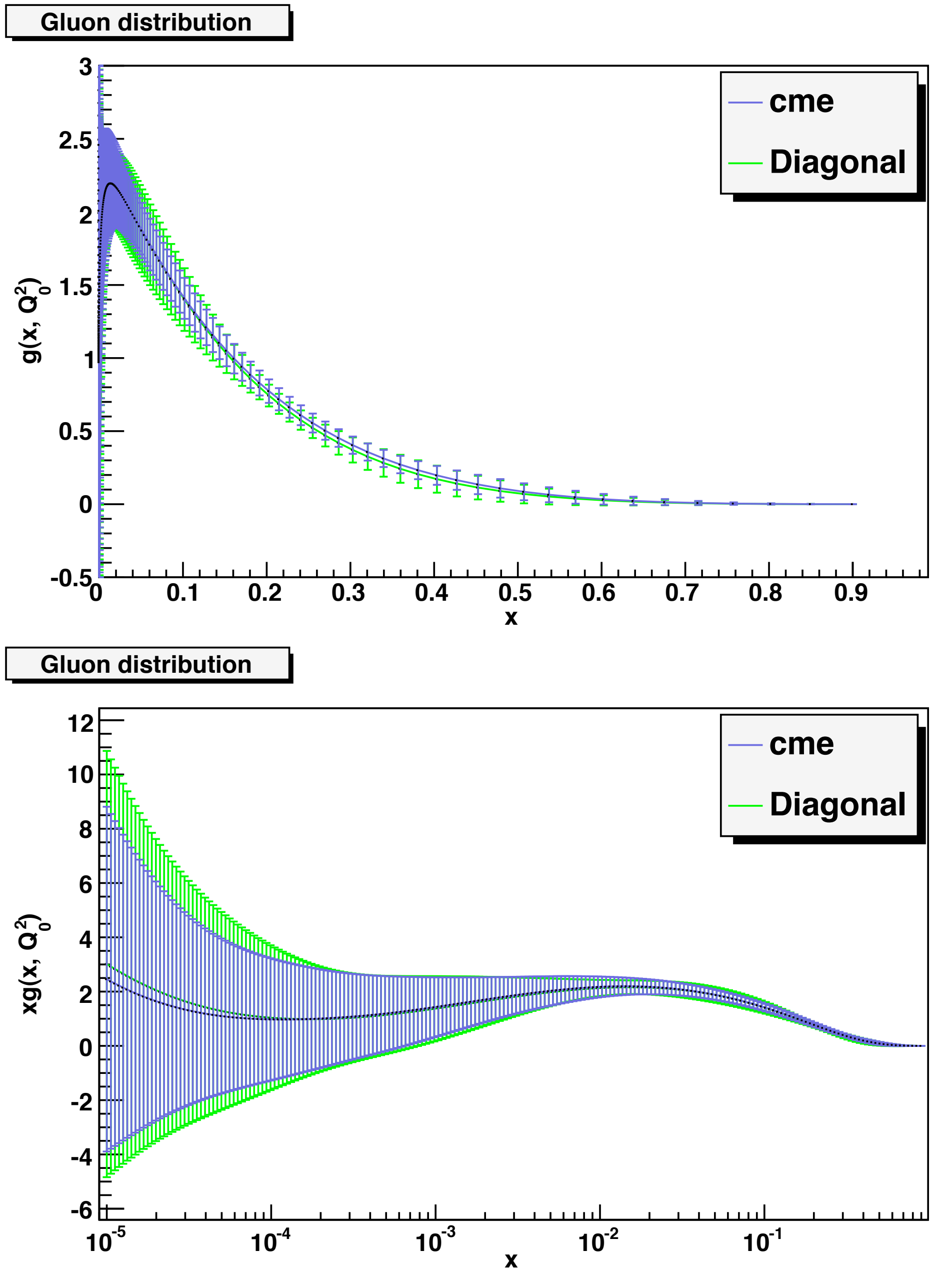

Figure 7: the gluon at the reference scale $Q_0^2 = 2\mathrm{GeV}^2$, plotted versus x on a linear (left) or a logaritmic scale (right).

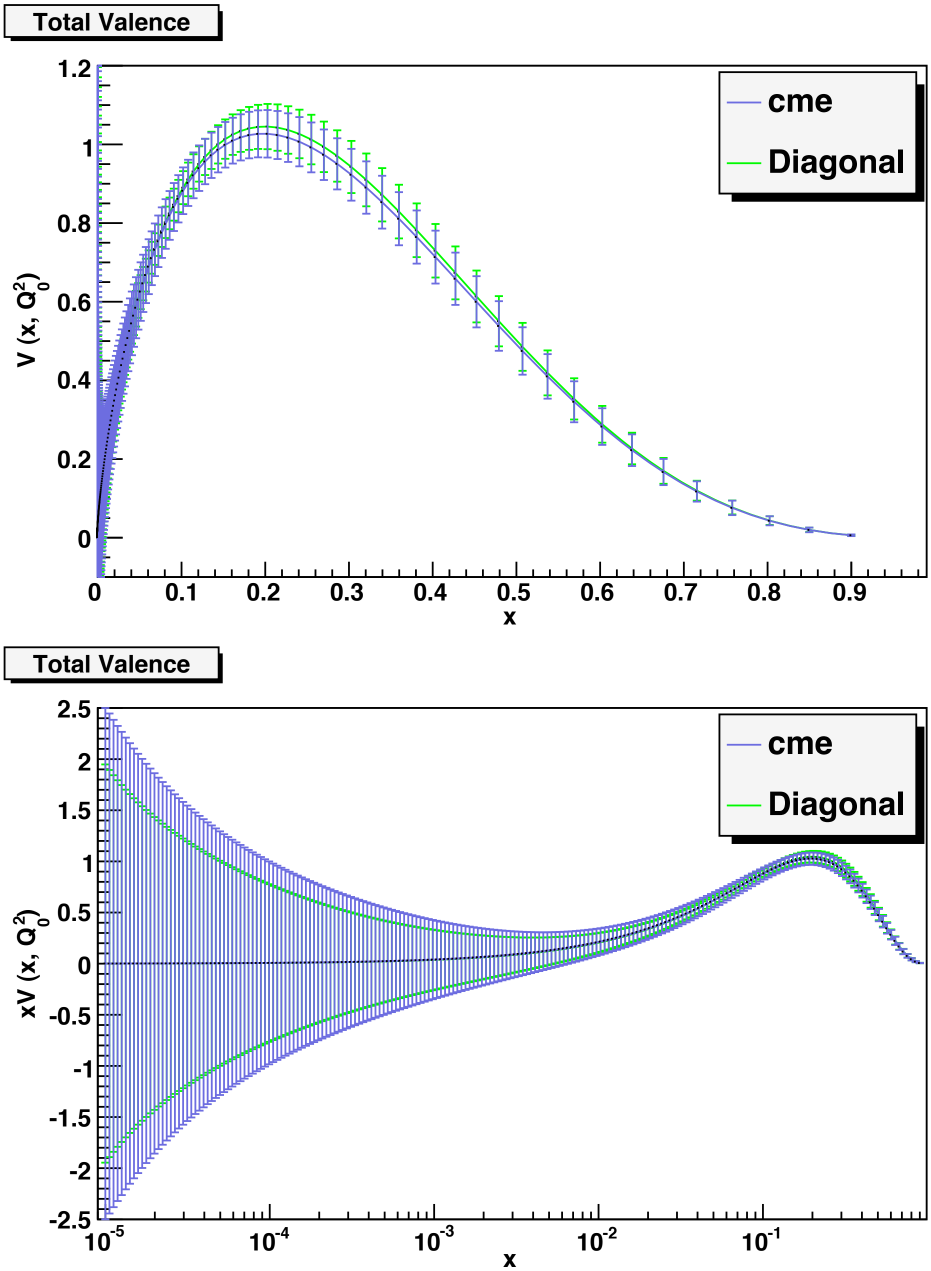

Figure 8: the total valence at the reference scale $Q_0^2 = 2\text{GeV}^2$, plotted versus x on a linear (left) or a logaritmic scale (right).

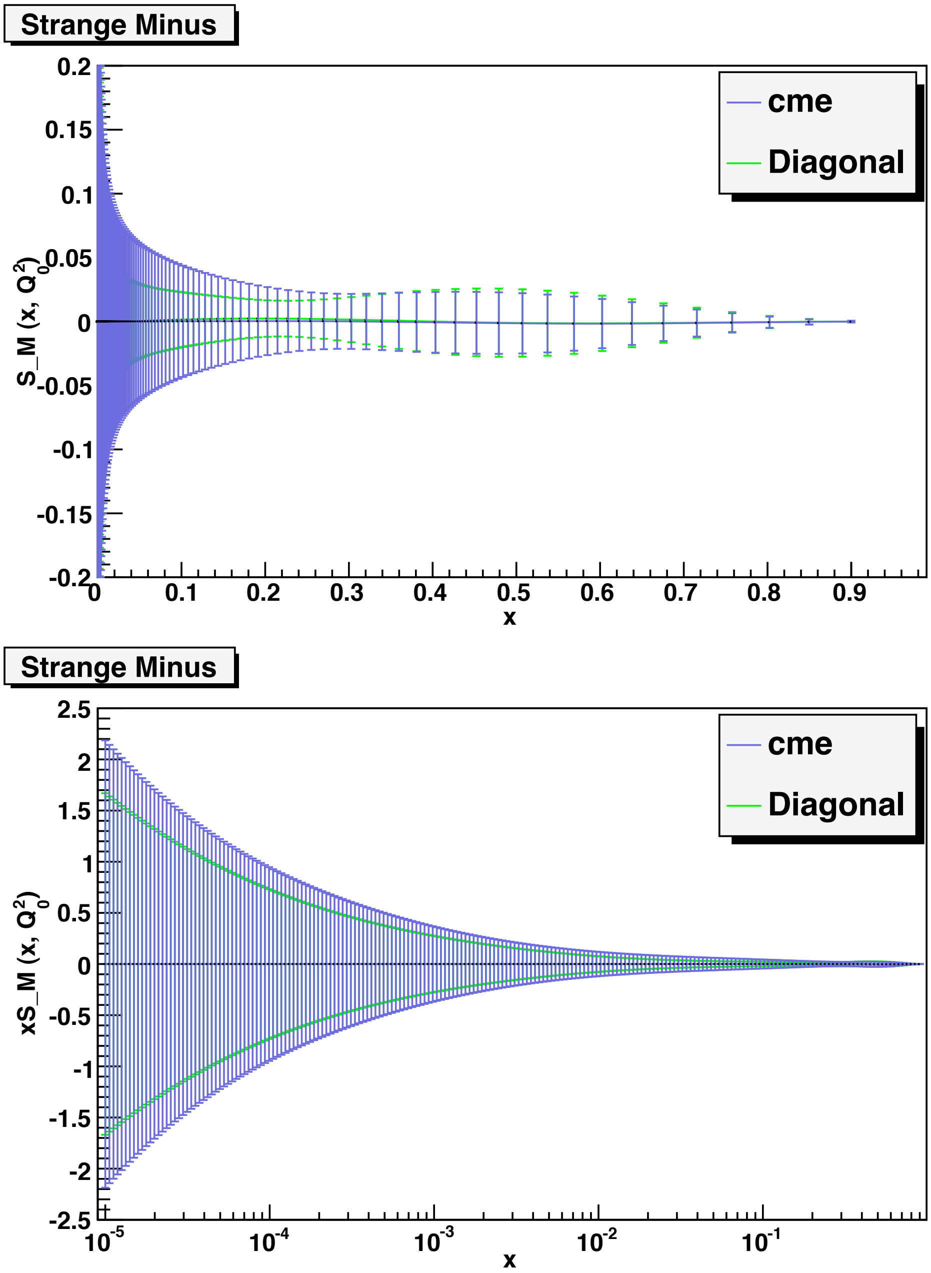

Figure 9: the strange minus at the reference scale $Q_0^2 = 2\mathrm{GeV}^2$, plotted versus x on a linear (left) or a logaritmic scale (right).

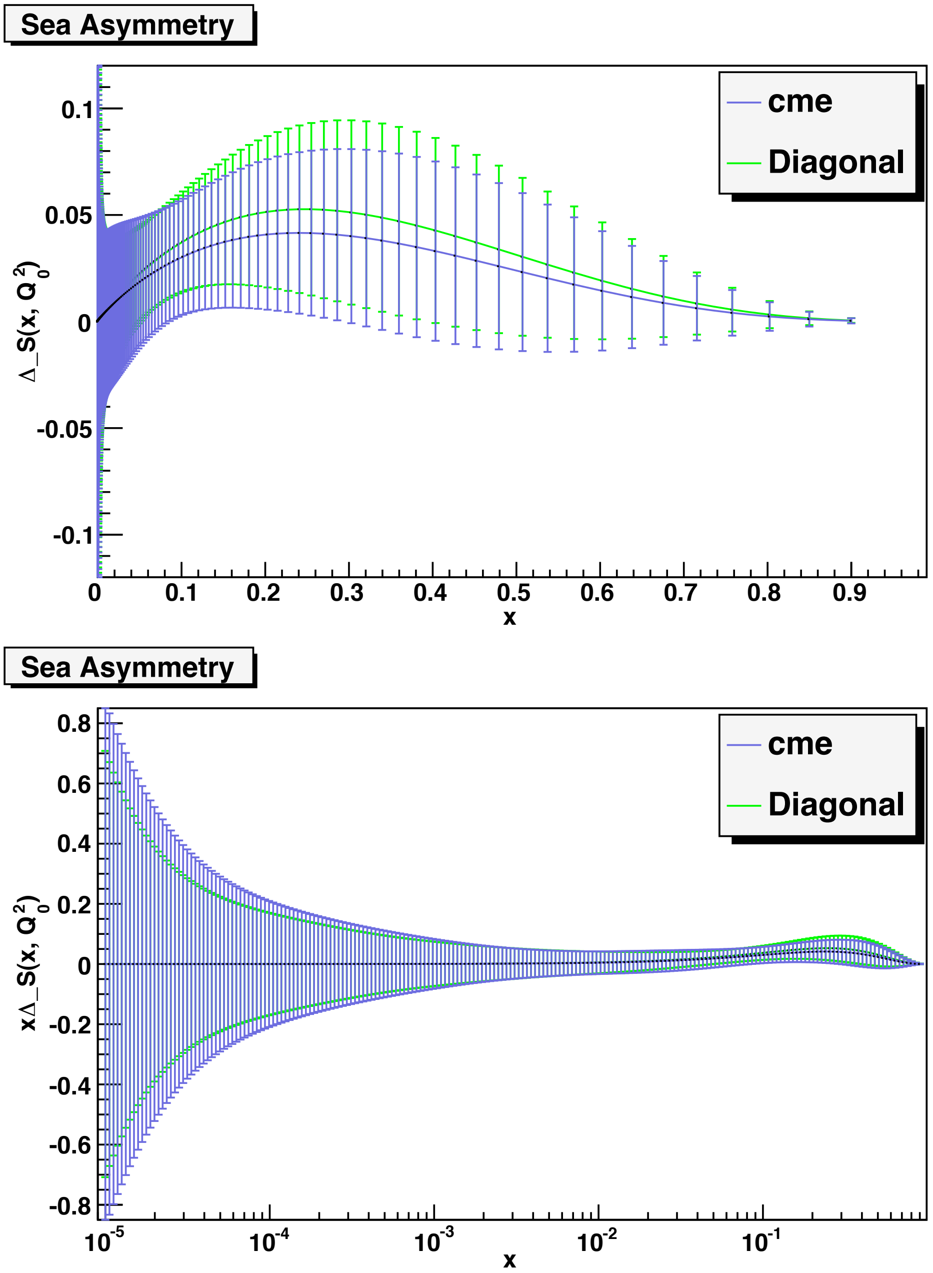

Figure 10: the sea asymmetry distribution at the reference scale $Q_0^2 = 2\mathrm{GeV}^2$, plotted versus x on a linear (left) or a logaritmic scale (right).

The gluon distribution shows a marked stability, as data come predominantly from HERA and ZEUS, which, as we have pointed out in sec. 3.1.1, have small systematic uncertainties. However, the distance is more pronounced in the region between $x = 10^{-5}$ and $x = 8 \cdot 10^{-4}$ (fig. 7), where data at low Q^2 are predominant: in this kinematical region, also ZEUS and HERA data have important correlations, as table 2 shows.

We conclude that the impact of the inclusion of correlations on a pdf global analysis is small, but not negligeble. Indeed, the averaged effect of correlations is small, but we have seen that locally differences between pdfs can be significant.
It would be interesting to determine quantitatively this effect, by calculating some observables for the LHC, e.g. the mass of the weak bosons $W^{\pm}$ and Z^0 for which the experimental accuracy is of about 1%. It is known that a very small difference in pdf uncertainties can be the cause of a considerable change for some observables.

3.2 Conclusions

We have studied the effect of the inclusion of correlated uncertainties on a global pdf analysis, by comparing a parton set obtained including correlations and a parton set produced by completely neglecting them.
Testing both the quality of the fit and the difference between pdfs, we have noticed that the inclusion of correlations do affect the pdf determination.
Namely, we see that, while the overall quality of the fit is essentially unchanged if measured with a χ^2 that contains the full covariance matrix, the purely diagonal χ^2 does change significantly. This is particularly noticeable for the subsets of data coming from experiments with large systematics. This suggests that a diagonal-χ^2-minimization artificially alters the weights of the various subsets of data, so that the experiments with large systematics are less important in the fit and viceversa. We conclude that the non-inclusion of correlations changes the fit, and not just its quality, which is confirmed by the fact that for some data sets with small systematics the quality of the two fits is very different.
Comparing the pdfs, we have noticed that the inclusion of correlations have a larger effect for the pdfs in the valence region, where the information is mostly controlled by the experiments with high systematics. The most significant differ-

ence is observed for the singlet and strange plus distributions: in some kinematical region it is of the order of the sum in quadrature of their standard deviations.
We conclude that, even if the effect of the inclusion of correlated systematics is small, it is not negligible, and for some data sets, and consequently some kinematical regions, it is quite significant.

References

[1] G. D. Coughlan, J. E. Dodd, *The Ideas of Particle Physics: an Introduction for Scientists*, Cambridge Univ. Press, 1991

[2] B. Povh, K. Rith, C. Scholz, F. Zetsche, *Particles and Nuclei*, Springer, 1995

[3] W. T. Giele, S. Keller, *Implications of Hadron Collider Observables on Parton Distribution Function Uncertainties*, Phys.Rev. D58 (1998) 094023, arXiv:hep-ph/9803393, 1998

[4] S. Catani *et al.*, *QCD*, arXiv:hep-ph/0005025, 2000

[5] S. Forte *et al.*, *Neural NetworkParametrizationof Deep–Inelastic Structure-Functions*, JHEP 0205 (2002) 062, arXiv:hep-ph/0204232, 2002

[6] M. Dittmar, S. Forte, A. Glazov, S. Moch, *Parton Distributions Summary Report for the HERA- LHC Workshop Proceedings*, arXiv:hep-ph/0511119, 2005

[7] S. Forte, *Structure Functions and Parton Distributions*, Nucl.Phys. A755 (2005) 100-110, arXiv:hep-ph/ 0502073, 2005

[8] NNPDF Collaboration, *Neural Network determination of parton distributions: the nonsinglet case*, JHEP 0703:039, arXiv:hep-ph/0701127, 2007

[9] S. Bethke, *Experimental Tests to Asymptotic Freedom*, Prog.Part.Nucl.Phys.58:351-386 arXiv:hep-ex/0606035, 2008

[10] A.D. Martin, W.J. Stirling, R.S. Thorne and G. Watt, *Parton Distributions for the LHC*, Eur.Phys.J.C63:189-285, arXiv:hep-ph/0901.0002, 2009

[11] Wu-Ki Tung, *Bjorken Scaling*, Scholarpedia, 2009

[12] NNPDF Collaboration, *A determination of parton distributions with faithful uncertainty estimation*, Nucl. Phys. B 809, arXiv:hep-ph/0808.1231, 2009

[13] NNPDF Collaboration, *Update on Neural Network Parton Distributions*, , arXiv:hep-ph/0811.2288, 2009

[14] NNPDF Collaboration, *Precision determination of electroweak parameters and the strange content of the proton from neutrino deep-inelastic scattering*, Nucl.Phys.B823:195-233, arXiv:hep-ph/0906.1958, 2009