UNIVERSITA' DEGLI STUDI DI MILANO
Facoltà di Scienze Matematiche, Fisiche e Naturali
Corso di laurea triennale in Fisica

Parton Distribution Function Reweighting in Perturbative QCD

Relatore: Prof. Stefano Forte
Correlatore: Dott. Stefano Carrazza

Laureando: Fabrizio Cimaglia
Matricola: 794224

Anno Accademico 2013-2014

"La fisica non è una rappresentazione della realtà, ma del nostro modo di pensare ad essa."

Werner Karl Heisenberg

Indice

Abstract

In questo elaborato analizziamo differenti tecniche usate per aggiungere l'informazione derivante da esperimenti ad un insieme di distribuzioni partoniche (PFDs).
Discutiamo il metodo del *Reweighting*, tramite il quale l'informazione contenuta in nuovi dati sperimentali viene aggiunta usando l'inferenza statistica Bayesiana assegnando un parametro, detto *peso*, ad ogni PDF. Analizziamo differenti espressioni per i pesi computazionalmente costruendo un modello generando dati pseudo-sperimentali distribuiti su una funzione armonica e fittando con differenti ensembles di PDFs; in particolare usiamo due forme funzionali per enfatizzare le discrepanze prodotte da diverse equazioni dei pesi. Mostriamo che un'espressione è più consistente con i dati sperimentali, mentre l'altra è solo valida in taluni casi. Infine implementiamo dei test statistici volti a confrontare gli indicatori di qualità dei Reweighting utilizzati nel modello.

1 Introduzione

1.1 Le Distribuzioni Partoniche in QCD

La fisica delle alte energie ai collider adronici dipende dalla conoscenza delle distribuzioni partoniche (PDFs), che contengono informazione sulla struttura interna degli adroni in termini di partoni; i partoni sono quarks e gluoni, i gradi di libertà dell'interazione forte.
Nella Cromo-Dinamica Quantistica (QCD), la teoria dell'interazione forte, le distribuzioni partoniche sono un input necessario per il calcolo di qualsiasi processo ai collider adronici e forniscono le più precise informazioni disponibili sulla struttura degli adroni. Dato un processo adronico, è necessario quindi disporre di una conoscenza adeguata delle PDFs. Attualmente, la determinazione delle PDFs non è banale: le distribuzioni partoniche non possono essere dedotte dai soli principi primi e anche la loro forma funzionale è ignota. Le PDFs devono quindi essere estratte da un insieme di dati sperimentali; questo è problematico in quanto i dati sperimentali sono insiemi discreti di punti e la determinazione di una funzione a partire da un numero finito di valori è un problema matematicamente mal posto.
Le PDFs vengono determinate supponendo una forma funzionale con un numero finito di parametri che possono essere calcolati fittando tali funzioni sui dati sperimentali. Solitamente, quando sono disponibili nuovi dati, le PDFs sono determinate fittando di nuovo tenendo conto della nuova informazione di cui si dispone. Tuttavia, vi è un modo alternativo per aggiornare un insieme di distribuzioni partoniche: generando le PDFs tramite metodi MonteCarlo (che verranno illustrati in seguito), il Teorema di Bayes può essere utilizzato per aggiornare la distribuzione di probabilità di partenza (prior), includendo l'informazione contenuta nei nuovi dati sperimentali. I metodi statistici, noti col nome *Reweighting*, che usano l'inferenza Bayesiana per aggiornare gli ensemble di distribuzioni partoniche hanno il vantaggio di essere del tutto liberi da ipotesi della forma funzionale, eccetto per la scelta del prior. Inoltre sono più efficienti in quanto usano solo i nuovi dati sperimentali assegnando un parametro ad ogni PDF, detto *peso*, che racchiude l'importanza di ogni funzione. Uno schema qualitativo di un metodo basato sull'inferenza Bayesiana è mostrato in Figura 1. Le PDFs sono tutte equiprobabili in quanto generate sui dati sperimentali tramite Importance Sampling: nel caso in cui siano disponibili nuovi datasets, il Teorema di Bayes permette di stimare ed aggiornare la probabilità di osservare una PDF f, dato un determinato valore del χ^2 sui nuovi dati, calcolando il peso da assegnare ad ogni funzione.

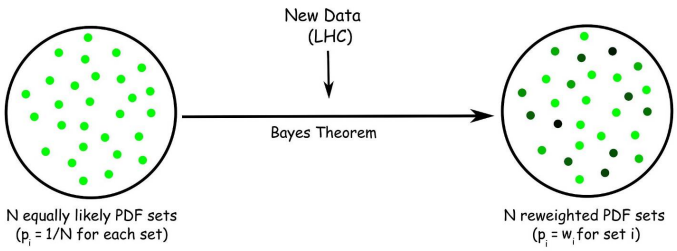

Figura 1: Una descrizione qualitativa del Reweighting: partendo da un set di PDF equiprobabili, aggiungendo nuovi dati sperimentali, ogni PDF guadagna un peso ω_i che rappresenta la probabilità condizionata di osservare un χ^2, data una PDF f.

L'obiettivo di questo elaborato è quello di esaminare i metodi che usano la statistica Bayesiana per aggiornare un set di distribuzioni partoniche. Svilupperemo un modello computazionale simulando settaggi sperimentali che rappresentano situazioni realistiche e discuteremo le prestazioni delle diverse espressioni dei pesi attualmente in uso implementando dei test statistici per confrontare alcuni indicatori di qualità degli ensembles utilizzati. Infine, i risultati computazionali verranno utilizzati per analizzare l'inferenza Bayesiana nelle distribuzioni partoniche da un punto di vista teorico.

L'elaborato è organizzato come segue: le prossime Sezioni si focalizzano sul problema della parametrizzazione delle distribuzioni partoniche, illustriamo in breve la scelta delle reti neurali come possibile forma funzionale. In seguito discuteremo del ruolo delle distribuzioni partoniche nelle fattorizzazioni delle sezioni d'urto per processi forti. Analizziamo alcuni esempi riguardo l'adro-produzione ed elettro-produzione in cui una corretta conoscenza delle PDFs permette di svolgere calcoli accurati sui valori di aspettazione delle osservabili fisiche. In seguito illustriamo l'equazione di evoluzione perturbativa a cui sono legate le PDFs a diversi ordini perturbativi in QCD.
La seconda Sezione di questo elaborato è dedicata allo stato dell'arte dei metodi noti per aggiornare ensembles di distribuzioni partoniche: in primo luogo illustriamo il metodo classico, detto Hessiano, tramite il quale si aggiorna un set di PDFs semplicemente aggiungendo il nuovo dataset a quello utilizzato per generare le PDFs per poi refittare l'ensemble di funzioni sulla combinazione dei due insiemi ottenuti tramite la loro unione. In secondo luogo sviluppiamo il formalismo del Reweighting che permette di assegnare un peso ad ogni PDF; infine discuteremo il problema del calcolo dei pesi applicando il formalismo introdotto a casi differenti.
Il terzo ed ultimo capitolo dell'elaborato riguarda la discussione dei risultati più significativi ottenuti tramite un modello computazionale: illustriamo il metodo usato per generare i dati sperimentali e l'algoritmo di minimizzazione implementato per generare gli ensembles di PDFs su cui implementiamo il Reweighting, quindi discutiamo i risultati più rilevanti tramite dei test statistici in modo da capire quale, fra i Reweighting utilizzati, fornisce risultati in miglior accordo con la teoria soggiacente.

1.2 Fattorizzazione delle PDFs

Le distribuzioni partoniche si usano nel calcolo di osservabili fisiche di processi di collisioni adroniche tramite fattorizzazione. La fattorizzazione delle sezioni d'urto è una proprietà fondamentale della QCD che permette di esprimere le sezioni d'urto tra adroni in termini di quelle relative alla collisione tra i loro costituenti. La sezione d'urto per un processo generico totale per la produzione di uno stato finale X in una collisione fra due adroni h_1 e h_2 può essere fattorizzata nella seguente forma

$$\sigma_X(s, M_X^2) = \sum_{a,b} \int_{x_{min}}^1 dx_1 dx_2 f_{a/h_1}(x_1, M_X^2) f_{b/h_2}(x_2, M_X^2) \tilde{\sigma}_{ab\to X}(x_1 x_2 s, M_X^2) \quad ,$$

$$= \sum_{a,b} \sigma_{ab}^0 \int_\tau^1 \frac{dx_1}{x_1} \int_{\tau/x_1}^1 \frac{dx_2}{x_2} f_{a/h_1}(x_1, M_X^2) f_{b/h_2}(x_2, M_X^2) C_{ab}\left(\frac{\tau}{x_1 x_2}, \alpha_S(M_X^2)\right) \quad ,$$

$$= \sum_{a,b} \sigma_{ab}^0 \int_\tau^1 \frac{dx}{x} \mathcal{L}_{ab}(x, M_X^2) C_{ab}\left(\frac{\tau}{x}, \alpha_S(M_X^2)\right) \tag{1}$$

dove s è l'energia del centro di massa della collisione adronica, il pre-fattore σ_{ab}^0 è la sezione d'urto al più basso ordine perturbativo sicché C è definita in modo tale che sia adimensionale e $\tilde{\sigma}_{ab\to X}(x_1 x_2 s, M_X^2)$ è la sezione d'urto partonica per la produzione dello stato finale X; il valore minimo di x_i è $\tau = x_{min}$, con

$$\tau \doteq \frac{M_X^2}{s}$$

che rappresenta la variabile di scaling del processo dove M_X^2 è la massa invariante dello stato finale. La sommatoria corre sui diversi tipi di partone, ossia: sei quarks, i corrispondenti antiquarks ed il gluone. In questo caso la funzione $f_{a/h_1}(x_i, M_X^2)$ è la funzione di distribuzione partonica dell'a-esimo partone nell'adrone incidente i-esimo. $\mathcal{L}_{ab}$ è la luminosità partonica definita come segue

$$\mathcal{L}_{ab} \doteq \int_x^1 \frac{dz}{z} f_{a/h_1}(z, M_X^2) f_{a/h_2}\left(\frac{x}{z}, M_X^2\right) \quad , \tag{2}$$

$$= \int_x^1 \frac{dz}{z} f_{a/h_1}\left(\frac{x}{z}, M_X^2\right) f_{a/h_2}(z, M_X^2), \tag{3}$$

dove definiamo z come segue

$$z \doteq \frac{\tau}{x_1 x_2}$$

Il coefficiente $C_{ab}(z, \alpha_S(M_X^2))$ è funzione della massa invariante M_X^2 e del rapporto adimensionale di M_X^2 diviso per l'energia del centro di massa $\widehat{s}$ del sotto-processo partonico

$$\frac{M_X^2}{\widehat{s}} = \frac{\tau}{x_1 x_2} \tag{4}$$

Nei casi più semplici, la sezione d'urto partonica al primo ordine è proporzionale ad una delta di Dirac di conservazione dell'energia: analoghi risultati fattorizzati possono essere ricavati per quantità più differenziali come, ad esempio, una distribuzione in rapidità soddisfacente le seguenti relazioni

$$\tilde{\sigma}_{ab\to X}(\tau, M_X^2) = \sigma_0 C_{ab}(\tau, \alpha_s(M_X^2)), \tag{5}$$

$$C_{ab}(x, \alpha_s(M_X^2)) = c_{ab}\delta(1-x) + \mathcal{O}(\alpha_s), \tag{6}$$

dove c_{ab} è una matrice con elementi che non si semplificano solo per gli stati di quark ed anti-quark.

Il risultato fattorizzato nell'Eq. (1) è valido sia per sezioni d'urto che per distribuzioni di rapidità

$$\frac{d\sigma}{dMX^2dY}(\tau, Y, M_X^2) = \sum_{ij} \int_{x_1^0}^1 dx_1 \int_{x_2^0}^1 dx_2 f_i^1(x_1, M_X^2) f_j^2(x_2, M_X^2) \times \frac{d\widehat{\sigma}_{ij}}{dMX^2dy}\left(\frac{\tau}{x_1x_2}, y, \alpha_s(M_X^2)\right), \tag{7}$$

dove la sezione d'urto adronica è differenziale rispetto alla rapidità Y dello stato finale X, mentre la sezione d'urto partonica è differenziale per la rapidità partonica y

$$y = Y - \frac{1}{2}\log\left(\frac{x_1}{x_2}\right); \tag{8}$$

le variabili x_1^0 e x_2^0 sono definite come segue

$$x_1^0 = \sqrt{\tau}e^Y \quad , \quad x_2^0 = \sqrt{\tau}e^{-Y}. \tag{9}$$

Dai casi di adro-fattorizzazione analizzati si evince la necessità di un'espressione esplicita delle distribuzioni partoniche per poter effettuare previsioni quantitative sui valori di aspettazione delle sezioni d'urto. Un altro aspetto importante della QCD che permette di legare le distribuzioni partoniche a diversi ordini perturbativi è rappresentato dall'equazione di evoluzione perturbativa.
L'Eq. (1) esprime la sezione d'urto in un processo adronico o adro-leptonico in termini delle distribuzioni partoniche alla scala M_X^2 del processo.
Le PDFs, a diverse scale, sono legate da un'equazione di evoluzione perturbativa

$$\frac{\partial}{\partial \log Q^2}\begin{pmatrix} \Sigma(x,Q^2) \\ g(x,Q^2) \end{pmatrix} = \int_x^1 \frac{dy}{y}\begin{pmatrix} P_{qq}^S(\frac{x}{y}, \alpha_S(Q^2)) & 2n_f P_{qg}^S(\frac{x}{y}, \alpha_S(Q^2)) \\ P_{gq}^S(\frac{x}{y}, \alpha_S(Q^2)) & P_{gg}^S(\frac{x}{y}, \alpha_S(Q^2)) \end{pmatrix}\begin{pmatrix} \Sigma(y,Q^2) \\ g(y,Q^2) \end{pmatrix}, \tag{10}$$

dove g è la distribuzione gluonica e Σ indica un singoletto di distribuzione di quark definito come

$$\Sigma(x,Q^2) \doteq \sum_{i=1}^{n_f} q_i(x,Q^2) + \overline{q}_i(x,Q^2). \tag{11}$$

Gli elementi di matrice P_{ij} sono serie perturbative in α_S.
In questo sviluppo perturbativo vi sono alcune condizioni dovute alle leggi di conservazione: in particolare la conservazione del numero barionico impone che

$$\int_0^1 dx \left[q_i(x,Q^2) + \overline{q}_i(x,Q^2)\right] = n_i \quad , \quad (n_u = 2, n_d = 1, n_{s,c,b,t} = 0), \tag{12}$$

dove gli indici u, s, c, b, t si riferiscono rispettivamente al quark up, strange, charm, bottom e top. La legge di conservazione del momento totale stabilisce che

$$\int_0^1 dx\, x\left\{\sum_{i=1}^{n_f}\left[q_i(x,Q^2) + \overline{q}_i(x,Q^2)\right] + g(x,Q^2)\right\} = 1. \tag{13}$$

Risolvendo l'Eq. (10) è possibile determinare le PDFs a qualunque scala in termini delle PDFs ad una qualsiasi scala di riferimento. Sostituendo il risultato dell'espansione nell'espressione fattorizzata per la sezione d'urto è possibile esprimere le sezioni d'urto a qualunque scala in termini delle PDFs ad una scala di riferimento.

1.3 Parametrizzazione delle PDFs

Un set di PDFS è un insieme di funzioni, una per ogni specie di partone. In principio vi sono 13 PDFs indipendenti di un adrone (6 per quarks ed anti-quarks ed una per il gluone). Solitamente le PDFs sono parametrizzate usando la seguente forma funzionale

$$f_i(x, Q^2) = x^{a_i}(1-x)^{b_i} g_i(x), \tag{14}$$

dove g_i è un polinomio in x o $\sqrt{x}$. Questa scelta garantisce che la PDF vada come una potenza di x per $x \to 0$ ed una potenza di $1-x$ per $x \to 0$. Solitamente, l'intero set di PDFs è parametrizzato con $20-30$ parametri liberi.

Un modo alternativo per parametrizzare le distribuzioni partoniche consiste nella scelta delle reti neurali (Neural Network (NN)), funzioni con un numero molto alto di parametri. Per definizione una rete neurale è un set di unità artificiali interconnesse chiamate *neuroni*; i neuroni sono costrutti matematici usati per modificare l'informazione in input.
Uno schema di una rete neurale è mostrato nella seguente Figura:

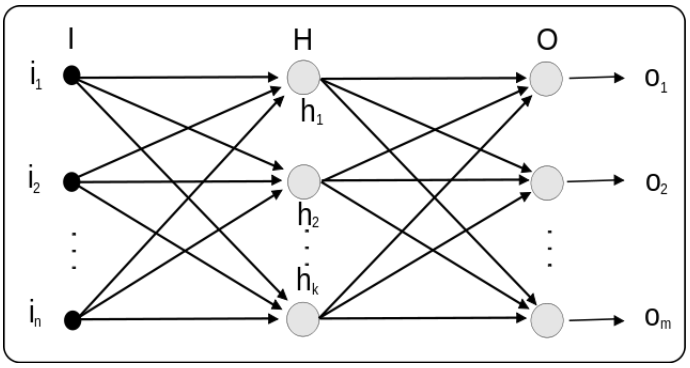

Figura 2: Questa Figura mostra una generica NN con i_n unità di input, h_k neuroni intermedi e O_m unità di output: l'informazione "viaggia" dalle celle iniziali, attraverso i neuroni modificandosi, fino alle celle finali. Ogni set di neuroni verticali è detto *layer* della rete neurale (la rete in figura è del tipo n-k-m).

Lo stato di attivazione di un neurone h_i è determinato come una funzione delle unità ad esso connesse; ogni coppia di unità (i, j) è collegata da sinapsi caratterizzate da un peso ω_{ij} e la funzione di attivazione è data da

$$\xi_i = g\left[\sum_j \omega_{ij}\xi_j - \theta_i\right], \tag{15}$$

dove ξ rappresenta lo stato di attivazione del neurone i, la somma su j scorre su tutti i neuroni connessi ad i e la funzione g è nota come funzione di attivazione; il numero di neuroni e layers interni definisce l'*Architettura* della rete neurale. La funzione g è solitamente non lineare, un esempio di g è dato dalla seguente relazione

$$g(x) = \frac{1}{1-e^{\beta x}} \tag{16}$$

Pur trattandosi, in ultima analisi di una parametrizzazione mediante funzioni non lineari, a tutti gli effetti pratici le reti neurali non richiedono la scelta di una forma funzionale esplicita. Poiché il numero di parametri è molto elevato, la rete neurale può riprodurre non solo la forma funzionale (ignota) soggiacente ai dati, ma anche le loro fluttuazioni statistiche. Questo è detto "overlearning". La necessità di evitare l'overlearning implica che il fit migliore non corrisponde al minimo del χ^2.

2 Metodi per aggiornare le PDFs

Questa Sezione è dedicata ad uno studio dei metodi ad oggi conosciuti per aggiornare un ensemble di distribuzioni partoniche $\varepsilon = \{f_k\}_{k=1}^{N}$.
Le distribuzioni partoniche sono funzioni, e come tali possono essere determinate da un numero finito di punti solo se si compiono delle assunzioni teoriche, ad esempio fissando una forma funzionale. Da qui in poi assumeremo che ogni PDF sia fittata su un set di dati sperimentali $\mathbf{y} \in \mathbb{R}^p$; le procedure attualmente note per aggiornare un esemble di PDFs sono dette *Metodo Hessiano* e *Reweighting*. Entrambi verranno discussi dettagliatamente nelle prossime Sezioni.
In generale, supponiamo di disporre un set di nuovi dati sperimentali $\mathbf{z} \in \mathbb{R}^n$: il metodo Hessiano stabilisce che l'informazione contenuta in $\mathbf{z}$ possa essere inclusa in ε semplicemente aggiungendo il nuovo vettore a quello utilizzato per fittare le PDF, quindi refittando le funzioni sul vettore $\mathbf{w} \doteq (\mathbf{y}, \mathbf{z}) \in \mathbb{R}^{p+n}$ dato dall'unione dei due. Il metodo del Reweighting stabilisce che $\mathbf{z}$ può essere aggiunto utilizzando l'inferenza Bayesiana [1] in termini di funzioni di densità di probabilità [2]. In accordo con questa guide-line, è sufficiente aggiornare ε tramite il solo utilizzo dei nuovi dati sperimentali e senza la necessità di un global-fit previsto dal metodo Hessiano.

2.1 Il metodo Hessiano

Supponiamo che siano dati un ensemble $\varepsilon = \{f_k\}_{k=1}^{N}$ di PDFs ed un vettore set sperimentali $(\mathbf{x}, \mathbf{y})$. Ogni f_k sarà una funzione $f(x, \mathbf{a})$ dipendente da una variabile x e da un set di parametri da determinarsi tramite un fit sui dati sperimentali. Supponiamo che sia noto il vettore $\mathbf{a}^0$ che permette di minimizzare un indicatore di bontà del fit come il χ^2, definito come segue

$$\chi^2(\mathbf{a}, \mathbf{y}) \doteq \sum_{ij} (y_i - f(x_i, \mathbf{a}))\Sigma_{ij}^{-1}(y_j - f(x_j, \mathbf{a})), \tag{17}$$

dove Σ_{ij} è la matrice di covarianza dei dati sperimentali. Definiamo la matrice Hessiana tramite la seguente equazione

$$H_{ij} \doteq \frac{1}{2}\frac{\partial^2 \chi^2}{\partial a_i \partial a_j}\left(\mathbf{a} = \mathbf{a}^0\right). \tag{18}$$

Il metodo Hessiano si focalizza su un espansione del χ^2 intorno al suo punto di minimo χ_0^2, da cui:

$$\chi^2 = \chi_0^2 + \sum_{ij}(a_i - a_i^0)H_{ij}(a_j - a_j^0) + \mathcal{O}(||\mathbf{a} - \mathbf{a}^0||^2). \tag{19}$$

La matrice H, per costruzione, è diagonalizzabile, con autovalori λ_i ed autovettori ψ_i. Per determinare uno sviluppo del χ^2 possiamo valutare la differenza $\mathbf{a} - \mathbf{a}^0$ tra un generico vettore di parametri e quello che determina il valore minimo del χ^2 in termini dei vettori $\{\psi_i\}_{i\in\Omega}$, dove Ω è un insieme di indici, come segue

$$\delta\mathbf{a} \doteq \mathbf{a} - \mathbf{a}^0 = \sum_k \xi_k \mathbf{e}_k \tag{20}$$

[1] **Teorema di Bayes**: Consideriamo uno Spazio d'eventi Ω, $\mathcal{I}$ un insieme di indici, $\{\mathcal{E}_i\}_{i=1}^{n}$ una partizione di Ω e $\mathcal{P}(A) > 0$ dove A è un evento in Ω. Per un generico elemento $\mathcal{E}_n$, con $n \in \mathcal{I}$, la probabilità condizionata che, osservato $\mathcal{E}_n$ la sua causa sia A, $\mathcal{P}(\mathcal{E}_n|A)$ è data da

$$\mathcal{P}(\mathcal{E}_n|A) = \frac{\mathcal{P}(A|\mathcal{E}_n)\mathcal{P}(\mathcal{E}_n)}{\sum_{i\in\mathcal{I}}\mathcal{P}(A|\mathcal{E}_i)\mathcal{P}(\mathcal{E}_i)}$$

■

[2] Per evitare confusione, useremo pdf per funzione di distribuzione di probabilità e PDF per funzione di distribuzione partonica.

dove il set $\mathbf{e}_k \doteq \sqrt{\lambda_k}\psi_k$ forma una base di uno spazio vettoriale reale n-dimensionale. Sostituendo l'Eq (20) nell'Eq. (19) otteniamo

$$\chi^2 \approx \chi_0^2 + \sum_{ij} \xi_i \xi_j (\mathbf{e}_i)^t H_{lm} \mathbf{e}_j \tag{21}$$

dove trascuriamo una quantità dell'ordine di $\mathcal{O}(||\mathbf{a} - \mathbf{a}^0||^2)$. Notando che $\sum_j (\mathbf{e}_i)^t H_{lm} \mathbf{e}_j = \delta_{ij}$, possiamo scrivere una forma finale dello sviluppo desiderato

$$\boxed{\chi^2(\mathbf{a}, \mathbf{y}) = \chi_0^2 + \sum_k \xi_k^2}. \tag{22}$$

2.2 Il Reweighting

Questa Sezione è dedicata ad uno studio del Reweighting; in primo luogo daremo una descrizione degli aspetti generali di questo metodo basato sull'inferenza statistica, in secondo luogo affronteremo il problema del calcolo dei pesi.

2.2.1 Aspetti generali

Come già citato, il formalismo del Reweighting si basa sull'applicazione Teorema di Bayes. Questo metodo si può applicare se è disponibile una rappresentazione Monte-Carlo della distribuzione di probabilità delle PDFs che chiamiamo $\mathcal{P}(f)$. Questo vuol dire che è disponibile un insieme di PDFs tale che il valore medio di un'osservabile $\mathcal{O}$ può essere calcolato come segue

$$\langle \mathcal{O} \rangle = \int \mathcal{O}[f]\mathcal{P}(f)\, Df = \frac{1}{N}\sum_{k=1}^{N} \mathcal{O}[f_k], \tag{23}$$

dove Df è una misura di integrazione sullo spazio delle PDFs e nell'ultimo passaggio abbiamo approssimato l'integrale con una somma discreta. Non discuteremo qui come questa rappresentazione possa essere ottenuta.

Consideriamo ora il caso in cui siano disponibili nuovi dati sperimentali $\mathbf{z} = (z_1, \ldots, z_n) \in \mathbb{R}^n$ ed assumiamo che siano statisticamente indipendenti da $\mathbf{y}$ [3].
Per riferirci al nostro dataset $\mathbf{z}$ usiamo un indicatore $\mathcal{D}$ che potrebbe essere, ad esempio, l'insieme dei χ^2 delle repliche su $\mathbf{z}$, oppure il vettore $\mathbf{z}$ stesso. Il Teorema di Bayes stabilisce che la probabilità condizionata $\mathcal{P}(f|\mathcal{D})$ [4] di osservare f, dato $\mathcal{D}$, si scrive come

$$\mathcal{P}(f|\mathcal{D}) = \frac{\mathcal{P}(\mathcal{D}|f)}{A_{\mathcal{D}}}\mathcal{P}(f), \tag{24}$$

dove $\mathcal{P}(\mathcal{D}|f)$ è la funzione di densità di probabilità di osservare $\mathcal{D}$, avendo osservato f e $\mathcal{A}_{\mathcal{D}}$ è il fattore di normalizzazione che compare nel Teorema di Bayes; la pdf $\mathcal{P}(f|\mathcal{D})$ viene interpretata come l'update del prior $\mathcal{P}(f)$.

Ricordando l'Eq. (24), il nuovo valore medio di $\mathcal{O}$ può essere calcolato come segue

$$\begin{aligned}
\langle \mathcal{O} \rangle_{new} &= \int \mathcal{O}[f]\mathcal{P}(f|\mathcal{D})\, Df, \\
&= \frac{1}{A_{\mathcal{D}}} \int \mathcal{O}[f]\mathcal{P}(\mathcal{D}|f)\mathcal{P}(f)\, Df, \\
&= \frac{1}{N A_{\mathcal{D}}} \sum_{k=1}^{N} \mathcal{P}(\mathcal{D}|f_k)\mathcal{O}[f_k].
\end{aligned} \tag{25}$$

Nell'Eq. (25) interpretiamo il pre-fattore

$$\omega(f_k) \doteq \frac{\mathcal{P}(\mathcal{D}|f_k)}{A_{\mathcal{D}}} \tag{26}$$

come il *peso* della PDF f_k calcolato sul dataset $\mathbf{z}$.

Il nuovo valore di aspettazione è quindi dato dalla seguente media pesata

$$\langle \mathcal{O} \rangle_{new} = \frac{1}{N}\sum_{k=1}^{N} \omega_k \mathcal{O}[f_k]. \tag{27}$$

[3] Questo garantisce la validità del prodotto logico di probabilità.
[4] Con la notazione $\mathcal{P}(A|B)$ intendiamo la probabilità di osservare A, dato B.

Si noti che l'Eq. (27) può essere applicata per ricavare una condizione di normalizzazione sui pesi ω_k, infatti richiedendo che $\mathcal{O}$ sia l'operatore unità $\mathcal{U}$ si ha che

$$\langle \mathcal{U} \rangle_{new} = 1 = \frac{1}{N} \sum_{k=1}^{N} \omega_k \mathcal{U}[f_k],$$

siccome $\mathcal{U}[f_k] = 1$, $\forall k = 1, ..., N$, abbiamo che

$$1 = \frac{1}{N\tilde{A_{\mathcal{D}}}} \sum_{k=1}^{N} \mathcal{P}(\mathcal{D}|f_k).$$

Otteniamo il seguente vincolo sul fattore di normalizzazione

$$\tilde{A_{\mathcal{D}}} = \frac{1}{N} \sum_{k=1}^{N} \mathcal{P}(\mathcal{D}|f_k). \tag{28}$$

Il vantaggio di questo metodo è chiaro: è possibile includere un nuovo dataset senza la necessità di un global fit, ma soltanto calcolando i pesi da assegnare ad ogni PDF sul dataset $\mathbf{z}$. Nelle seguenti Sezioni mostreremo due possibili espressioni dei pesi ω_k, in primo luogo usando $\mathcal{D} = \chi^2$, in seguito con $\mathcal{D} = \mathbf{z}$.

2.2.2 L'equazione dei pesi di Giele-Keller

Un possibile modo per dedurre un'equazione dei pesi si basa sull'integrazione della distribuzione di probabilità $\mathcal{P}(f|\mathbf{z})$ sullo spazio dei dati sperimentali $\mathbf{z}$; in particolare domandandosi quale sia la probabilità per $\mathbf{z}$ di essere confinato in un volume differenziale $d^n\mathbf{z}$.
Il Teorema di Bayes può essere enunciato in termini di pdf come segue

$$\mathcal{P}(f|\mathbf{z})\mathcal{D}f\mathcal{P}(\mathbf{z})d^n\mathbf{z} = \mathcal{P}(\mathbf{z}|f)d^n\mathbf{z}\mathcal{P}(f)\mathcal{D}f, \tag{29}$$

dove $\mathcal{D}f$ è la misura di integrazione sullo spazio dei dati sperimentali, $\mathcal{P}(f)$ è la densità di probabilità rappresentata dall'ensemble ε, $\mathcal{P}(f|\mathbf{z})$ è l'update del prior $\mathcal{P}(f)$, dato $\mathbf{z}$. $\mathcal{P}(\mathbf{z})$ è la pdf prior nello spazio dei dati; la sola richiesta che facciamo su quest'ultima pdf è che non sia dipendente dalle PDFs f_k di ε. Per definire $\mathcal{P}(f|\mathbf{z})$, possiamo integrare l'Eq. (29) su una piccola sfera S_δ di raggio δ centrata su $\mathbf{z}$. L'espressione di sinistra dell'Eq. (29) diventa

$$\int_{S_\varepsilon} \mathcal{P}(f|\mathbf{z})\mathcal{D}f\mathcal{P}(\mathbf{z})d^n\mathbf{z} = \left[\frac{\delta^n}{n}\Omega_n\mathcal{P}(\mathbf{z})\right]\mathcal{P}(f|\mathbf{z})\mathcal{D}f, \tag{30}$$

dove Ω_n è l'angolo solido n-dimensionale. Integrando l'espressione di destra in maniera simile, i fattori di volume $d^n\mathbf{z}$ si elidono in entrambi i lati, quindi possiamo considerare il limite $\delta \to 0$ per ottenere

$$\mathcal{P}(f|\mathbf{z})\mathcal{D}f = \frac{\mathcal{P}(\mathbf{z}|f)}{\mathcal{P}(\mathbf{z})}\mathcal{P}(f)\mathcal{D}f. \tag{31}$$

Si noti che $\mathcal{P}(\mathbf{z}|f)$ è la funzione di verosimiglianza dei dati $\mathbf{z}$: assumiamo che i dati abbiano degli errori gaussiani dati da una matrice di covarianza Σ, centrati sulla previsione teorica $\mathbf{z}[f]$ [5].
Definendo il $\chi^2(\mathbf{z}|f)$ come

$$\chi^2(\mathbf{z}|f) \doteq (\mathbf{z} - \mathbf{z}[f])\Sigma^{-1}(\mathbf{z} - \mathbf{z}[f]), \tag{32}$$

segue che $\mathcal{P}(\mathbf{z}|f)$ è data da

[5] Con questa notazione intendiamo che la coppia di dati sperimentali (x_i, z_i) ha una previsione teorica soggiacente $(x_i, z_i[f])$, dove abbiamo identificato $z_i[f] \doteq f(x_i)$.

$$\mathcal{P}(\mathbf{z}|f)\mathcal{D}f \propto e^{-\frac{1}{2}\chi^2(\mathbf{z}|f)}\mathcal{P}(f)\mathcal{D}f. \tag{33}$$

Le Eqs. (33) e (28) ci conducono all'equazione dei pesi di Giele-Keller

$$\boxed{\omega_k = \frac{e^{-\frac{1}{2}\chi_k^2}}{\frac{1}{N}\sum_{j=1}^{N} e^{-\frac{1}{2}\chi_j^2}}}. \tag{34}$$

La deduzione dell'Eq. (34) si basa sulla possibilità di elidere gli elementi di volume $d^n\mathbf{z}$ da entrambi i membri dell'Eq. (29) dopo aver integrato su di una sfera di raggio δ. Il limite che stiamo considerando è quindi $d^n\mathbf{z} \to \mathbf{0}$. Tuttavia, l'elemento di volume $d^n\mathbf{z}$ appartiene ad uno spazio vettoriale reale n-dimensionale ed il modo in cui $d^n\mathbf{z}$ tende a zero può determinare, in generale, diversi risultati di questo limite; nel nostro caso, prendendo $d^n\mathbf{z} \to \mathbf{0}$ stiamo selezionando in modo univoco il vettore $\mathbf{z}$ senza tenere conto del fatto che nello spazio vettoriale dei dati sperimentali possano esistere diversi vettori caratterizzati dallo stesso valore del χ^2 per una data PDF. Vedremo nella prossima Sezione come poter tenere conto di questo aspetto nel passaggio in cui si integra l'Eq. (29) .

2.2.3 L'equazione dei pesi NNPDF

Un modo differente da quello appena illustrato per dedurre un'equazione dei pesi è dato dal metodo NNPDF. Supponiamo che il nuovo dataset $\mathbf{z}$ sia indipendente da quello utilizzato per fittare le PDFs. Si noti che quando fittiamo le PDFs sui dati sperimentali, non richiediamo che la previsione teorica $f(x_i)$ di un dato sperimentale y_i coincida con quest'ultimo, bensì che un indicatore di bontà del fit come il $\chi^2(\mathbf{z}|f)$ sia ottimizzato. Quindi, per dedurre un'equazione di Reweighting dobbiamo integrare su tutto $\mathbf{z}$ con la sola richiesta che $\chi^2(\mathbf{z}|f) = \chi^2$, per qualche valore fissato χ^2. Per semplicità utilizzeremo $\chi \doteq \sqrt{\chi^2(\mathbf{z}|f)}$ invece di χ^2 in modo tale da poter interpretare χ come la coordinata radiale in un sistema di coordinate polari-sferiche nello spazio di funzioni, centrato sul valore teorico $f(x_i)$ (per il dato z_i). L'espressione di sinistra del Teorema formulato come in Eq. (29) diventa

$$\int \delta(\chi - \chi(\mathbf{z}'|f))\mathcal{P}(f|\mathbf{z}')\mathcal{D}f\mathcal{P}(\mathbf{z}')d^n\mathbf{z} \propto \mathcal{P}(f|\chi)\mathcal{D}f, \tag{35}$$

definendo $\mathcal{P}(f|\chi)$ come una costante globale, possiamo stimare il suo valore tramite un'integrazione simile nell'espressione di destra dell'Eq. (29):

$$\int \delta(\chi - \chi(\mathbf{z}'|f))\mathcal{P}(\mathbf{z}'|f)\mathcal{D}f\mathcal{P}(\mathbf{z}')d^n\mathbf{z} = \frac{2^{1-\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}\Omega_n\chi^{n-1}e^{-\chi^2/2}\mathcal{P}(f)\mathcal{D}f, \tag{36}$$

dove abbiamo stimato l'integrale usando coordinate polari-sferiche, Ω_n è l'angolo solido n-dimensionale e $\Gamma\left(\frac{n}{2}\right)$ è la funzione Gamma di Eulero. Confrontando l'Eq. (35) e l'Eq. (36) si ha un'espressione esplicita della densità di probabilità $\mathcal{P}(f|\chi)$

$$\mathcal{P}(f|\chi)\mathcal{D}f \propto (\chi^2)^{\frac{n-1}{2}}e^{-\chi^2/2}\mathcal{P}(f)\mathcal{D}f. \tag{37}$$

Per dedurre un'equazione dei pesi da associare ad ogni PDF, è necessario definire la probabilità per ogni f integrando su una regione finita di spazio. Consideriamo l'intervallo $\chi_k \leq \chi \leq \chi_k + \delta$:

$$\int_{\chi_k}^{\chi_k+\delta} d\chi\mathcal{P}(f_k|\chi) = \delta \cdot \mathcal{P}(f_k|\chi_k), \tag{38}$$

dove assumiamo che $\delta \ll 1$. L'equazione precedente corrisponde all'Eq. (35) integrata su di una regione di spazio di spessore δ. Si noti che δ deve essere indipendente dalla scelta della replica f, altrimenti il risultato sarebbe errato e valido solo per una PDF.

Usando l'Eq. (38) e (37) otteniamo una relazione di proporzionalità fra i pesi ω_k e la pdf $\mathcal{P}(f_k|\chi_k)$

$$\omega_k \propto \mathcal{P}(f_k|\chi_k) \propto (\chi^2)^{\frac{n-1}{2}}e^{-\chi^2/2}. \tag{39}$$

Il fattore di normalizzazione è dato dall'applicazione della relazione (28) in (39) tramite i seguenti passaggi

$$\boxed{\omega_k = \frac{(\chi^2_k)^{\frac{n-1}{2}} e^{-\frac{1}{2}\chi^2_k}}{\frac{1}{N}\sum_{j=1}^{N}(\chi^2_j)^{\frac{n-1}{2}} e^{-\frac{1}{2}\chi^2_j}}}. \tag{40}$$

Si noti che quando $n \neq 1$ il risultato è chiaramente differente dall'equazione di Giele-Keller. L'andamento delle due equazioni dei pesi è mostrato in Figura 3.

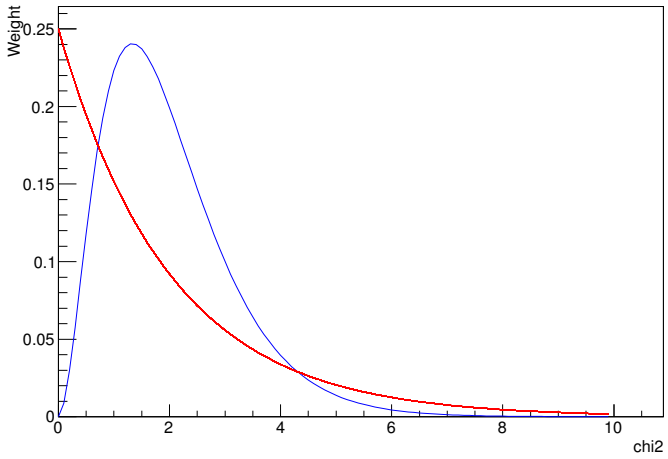

Figure 3: Gli andamenti delle equazioni dei pesi NNPDF ($n = 2$) e di Giele-Keller; rispettivamente Eq. (40) e (34).

La presenza del fattore $(\chi^2_k)^{\frac{n-1}{2}}$ in Eq. (40) determina che quando sono disponibili molti dati sperimentali, più i χ^2_k aumentano, piu il peso ω_k diminuisce; allo stesso modo valori molto piccoli dei χ^2_k determinano un peso sempre minore, cosa che non accade nell'Eq. (34) che privilegia valori del χ^2 sempre minori. Un altro aspetto importante nella deduzione NNPDF riguarda la scelta del volume di integrazione; la nostra scelta include tutti i punti nello spazio dei dati sperimentali con un particolare χ^2, ed una regione di spazio con uno spessore δ indipendente dal raggio $\chi(\mathbf{z}|f)$, nello stesso modo in cui nell'Eq. (30) il raggio della sfera è indipendente dal valore centrale. In entrambi i casi la giustificazione è che la misura di integrazione $d^n\mathbf{z}$ sullo spazio dei dati sperimentali è uniforme, nel senso che volumi uguali hanno uguale probabilità [6].

[6] Si noti che questa assunzione deve essere rispettata, altrimenti la pdf $\mathcal{P}(\mathbf{z}|f)$ non sarebbe una gaussiana.

2.3 Il paradosso di Borel-Kolmogorov

La ragione per cui le equazioni NNPDF e di Giele-Keller sono differenti è il paradosso di Borel-Kolmogorov. Quando consideriamo delle distribuzioni di probabilità multidimensionali è necessario prestare cautela agli insiemi di misura nulla poichè le probabilità condizionalte non sono ivi ben definite. Nei casi di interesse per le PDFs consideriamo due limiti differenti: nel caso NNPDF ci chiediamo quale sia la probabilità che il χ^2 di una replica appartenga all'intervallo $[\chi, \chi + d\chi]$, quindi il limite in considerazione è $d\chi \to 0$. Ovviamente $\chi^2 \in \mathbb{R}$ ed in uno spazio 1-dimensionale non vi sono alcune ambiguità sulle probabilità condizionate.
Nel caso Giele-Keller la distribuzione di probabilità $\mathcal{P}(f|\mathbf{z})$ appartiene ad uno spazio reale n-dimensionale, quindi in un volume $\tau_n = d^n\mathbf{z}$; in questo caso il limite da prendere in considerazione è $\tau_n \to \mathbf{0}$, ma in $\mathbb{R}^n$ la probabilità condizionata non è ben definita in quanto dipende dal modo in cui τ_n tende a zero. Il modo in cui $\tau_n \to \mathbf{0}$ implica che il vettore $\mathbf{z}$ sia univocamente selezionato senza curarsi del fatto che più di un solo vettore abbia lo stesso valore di χ^2, e quindi lo stesso risultato per una PDF f. Dobbiamo quindi includere tutti i vettori con lo stesso peso quando calcoliamo un'equazione di Reweighting e sommare su tutti i volumi τ_n che formano la superficie di $\chi(\mathbf{z}|f)$. In questo modo il volume τ_n diventa un guscio sferico con raggio $d\chi$ ed il limite $\tau_n \to \mathbf{0}$ può essere considerato senza ambiguità.

La Sezione seguente si focalizza sui più importanti risultati di questo elaborato. Verificheremo il funzionamento delle differenti espressioni dei pesi discusse nelle Sezioni 2.2.2 e 2.2.3 tramite un modello computazionale.

3 Risultati computazionali

In questa Sezione verificheremo le espressioni dei pesi analizzate nel capitolo precedente tramite un modello numerico; i risultati più significativi verranno discussi attraverso dei test volti a confrontare alcuni indicatori statistici di qualità degli ensembles utilizzati.

3.1 Descrizione del modello

Costruiremo un modello generando i dati sperimentali a partire da una funzione armonica aggiungendo ad ogni punto una perturbazione gaussiana; il generico dato sperimentale sarà quindi della forma (x_i, y_i), dove $y_i = \sin(x_i) + r_i$, ed r_i è la i-esima perturbazione gaussiana. In seguito fitteremo i dati sperimentali con due forme funzionali per poi implementare l'algoritmo del Reweighting su un nuovo set di dati sperimentali. Infine, discuteremo i risultati computazionali tramite un test statistico suddiviso in due parti: la prima in cui confrontiamo il χ^2 delle PDFs ottenute come medie pesate, con pesi dati dalle equazioni NNPDF e di Giele-Keller, delle PDFs che formano un ensemble ε con quello della funzione soggiacente ai dati sperimentali; la seconda in cui confrontiamo le incertezze delle PDFs.

3.1.1 Simulazione dei dati sperimentali

I dati sperimentali sono generati a partire dalla seguente funzione

$$f(x, \mathbf{a}) = a_1 \sin(a_2 x) \quad , \quad \mathbf{a} \doteq (a_1, a_2) \tag{41}$$

aggiungendo delle perturbazioni gaussiane ad ogni valore di f. Il vettore $\mathbf{a}$ è scelto convenzionalmente come $\mathbf{a} = (1, 1)$. In generale quindi, il set di dati sperimentali sarà del tipo $(\mathbf{x}, \mathbf{y}) \doteq (x_1, y_1; \ldots; x_n, y_n) = (x_1, \sin(x_1) + r_1; \ldots; x_n, \sin(x_n) + r_n)$, con $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$; ad ogni y_i è associato un parametro σ_i che rappresenta l'incertezza sperimentale sul dato i-esimo. Dataset differenti sono ottenuti variando l'ampiezza della gaussiana utilizzata per generare le perturbazioni r_i.

3.1.2 Generazione delle PDFs

Utilizziamo le seguenti forme funzionali per fittare i dati sperimentali generati come illustrato nella precedente Sezione :

- **Ensemble sinusoidale**:

$$f(x) = \alpha \sin(\beta x) + \gamma \cos(\delta x). \tag{42}$$

  Useremo questo ensemble per studiare i possibili risultati di un fit effettuato con una forma funzionale simile a quella utilizzata per generare i dati sperimentali.

- **Ensemble polinomiale**:

$$f(x) = \sum_{i=1}^{p} a_i x^i, \tag{43}$$

  dove p è il grado del polinomio. In un esperimento reale, la funzione soggiacente ai dati sperimentali è ignota, da cui la necessità di un fit con un ensemble diverso dall'Eq. (41).

Il problema che analizziamo ora riguarda la generazione di un ensemble di PDFs tramite un metodo MonteCarlo:

L'algoritmo di minimizzazione. Supponiamo di aver scelto una delle forme funzionali descritte sopra ed un insieme di dati sperimentali $(\mathbf{x}, \mathbf{y})$. La generazione di un set di PDFs può essere realizzata tramite un algoritmo genetico basato sulla minimizzazione del χ^2 di ogni funzione, iterando un numero finito di volte.
In particolare, partiamo da uno stato iniziale costituito da N mutanti [7] (ad esempio $N = 100$) in cui fissiamo i parametri delle funzioni con dei numeri distribuiti su una gaussiana centrata in zero. Calcoliamo il χ^2 di ogni mutante sui dati $\mathbf{y}$ e selezioniamo la funzione con minor χ^2 come miglior funzione: f_{best}. In seguito copiamo i parametri di f_{best} su tutti gli altri mutanti e, per ogni mutante, cambiamo un solo parametro usando numeri distribuiti sulla stessa gaussiana utilizzata per lo stato iniziale. Ricalcoliamo il χ^2 di ogni funzione sui dati sperimentali ed iteriamo questa procedura N_{it} volte. Al termine delle N_{it} iterazioni, il mutante con χ^2 minore diventerà la prima PDF dell'ensemble. Per ottenere le altre PDFs cambiamo il valore iniziale che fissa lo stato iniziale e ripetiamo la procedura di minimizzazione. Questo algoritmo viene ripetuto N volte, dove N è il numero di PDFs che formano l'ensemble su cui implementeremo il Reweighting.

Si noti che l'algoritmo usato presenta dei limiti: trattandosi di un algoritmo a lunghezza fissata (numero finito di iterazioni) è necessario verificare la scelta del numero di iterazioni in relazione alla forma funzionale scelta, in modo da accertarsi che il χ^2 raggiunga un valore minimo sufficientemente vicino a quello teorico, che tende al numero di dati sperimentali. In generale, può essere necessario un numero diverso di iterazioni fra l'ensemble sinusoidale e quello polinomiale se disponiamo di polinomi con un grado abbastanza elevato; infatti, in base alla gaussiana usata per assegnare i parametri, un polinomio può crescere in modulo molto rapidamente, rovinando la qualità del fit e dell'ensemble. E' necessario quindi trovare la "giusta" combinazione di parametri per fittare un dato set-up sperimentale. Un possibile modo per accertarsi di aver iterato l'algoritmo di minimizzazione un numero sufficientemente alto di volte consiste nel plottare, ad ogni iterazione i-esima, con $i = 1, \ldots, N_{it}$, il χ^2 di f_{best} in funzione del passo d'iterazione; in questo modo l'andamento dei punti deve essere decrescente all'aumentare dell'indice di iterazione. Per verificare di aver raggiunto un valore "stabile" del χ^2 (vicino al valore teorico) possiamo ripetere il plot utilizzando un numero di iterazioni ad esempio pari a $3N_{it}$; se N_{it} è stato scelto correttamente, l'andamento del χ^2 per $i > N_{it}$ dovrebbe essere costante. Riportiamo di seguito i grafici ottenuti per il calcolo della prima PDF dell'ensemble sinusoidale e polinomiale (con grado pari a 4) definiti precedentemente, con 100 dati sperimentali ed incertezze pari a 0.5; i grafici sono ottenuti in scala log-log per mettere in evidenza le regioni in cui il χ^2 è stabile.

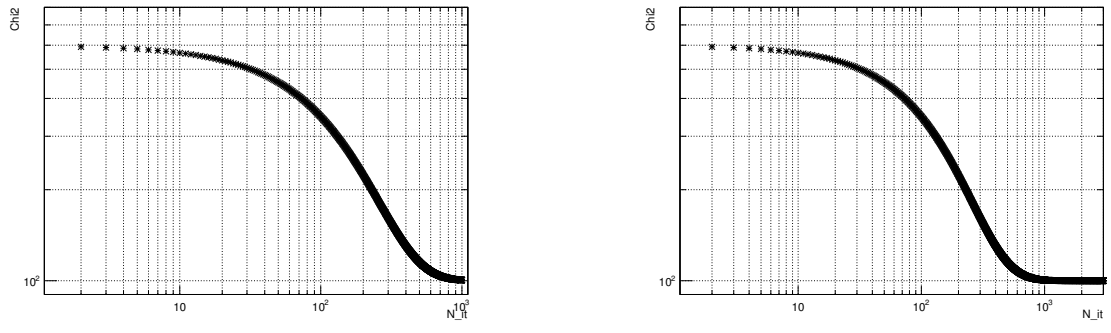

Figura 4: A sinistra l'andamento del χ^2 della prima PDF sinusoidale in funzione del numero di iterazioni $N_{it} = 1000$; si nota che i punti tendono al numero di dati sperimentali utilizzati $n = 100$ ma, con 1000 iterazioni, siamo all'inizio dell'intervallo di stabilità del χ^2. A destra l'andamento degli stessi punti; con un numero di iterazioni pari a 3000, si nota che il valore del χ^2 nell'intervallo $i \in [1000, 3000]$ è costante.

[7] Col termine *mutante* ci riferiamo ad una funzione $f(x, \mathbf{a})$ che durante lo svolgimento dell'algoritmo cambia il set di parametri $\mathbf{a}$ ottimizzando il χ^2 di f sui dati sperimentali.

Analogamente a quanto fatto per l'ensemble sinusoidale, riportiamo gli andamenti del χ^2 per l'ensemble polinomiale con lo stesso settaggio sperimentale.

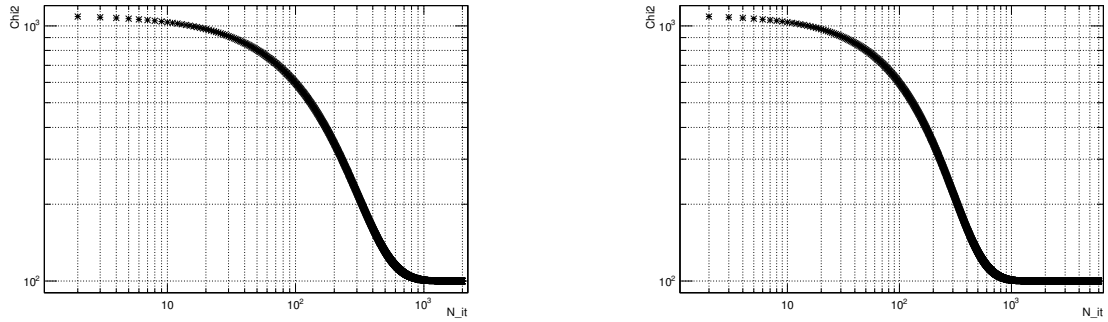

Figura 5: A sinistra l'andamento del χ^2 della prima PDF polinomiale in funzione del numero di iterazioni $N_{it} = 2000$; anche in questo caso si nota che i punti tendono al numero di dati sperimentali utilizzati $n = 100$. A destra l'andamento degli stessi punti; con un numero di iterazioni pari a 6000. Come in Figura 4, il valore del χ^2 nell'intervallo $i \in [2000, 6000]$ è costante. A parità di settaggio sperimentale, con l'ensemble polinomiale è necessario iterare l'algoritmo un numero maggiore di volte per raggiungere un valore costante del χ^2 rispetto all'ensemble sinusoidale; come si nota dall'andamento del grafici in un intorno di $i = 1000$, il χ^2 non è ancora stabile.

I grafici mostrati indicano che sia per l'ensemble sinusoidale che per quello polinomiale la scelta della lunghezza d'iterazione è idonea per la stabilizzazione del χ^2; tuttavia, per verificare le effettive prestazioni dell'algoritmo su diversi indicatori di qualità del fit che verranno discussi nella Sezione 3.1.3, implementeremo il metodo Hessiano ed il Reweighting usando diversi valori di N_{it} come $N_{it} = 300, 600, 1000, 2000, 3000$ nel caso sinusoidale e $N_{it} = 600, 1200, 2000, 4000, 6000$ per l'ensemble polinomiale. Per ogni valore di N_{it} discuteremo i risultati ottenuti. Usando $N_{it} = 1000$ per l'insieme sinusoidale e $N_{it} = 2000$ per quello polinomiale, mostriamo di seguito le distribuzioni dei χ^2 ridotti delle repliche[8]. I valori dei χ^2 sono calcolati sui 100 dati sperimentali usati per fittare le PDFs durante l'esecuzione dell'algoritmo di minimizzazione, prima dell'implementazione del Reweighting.

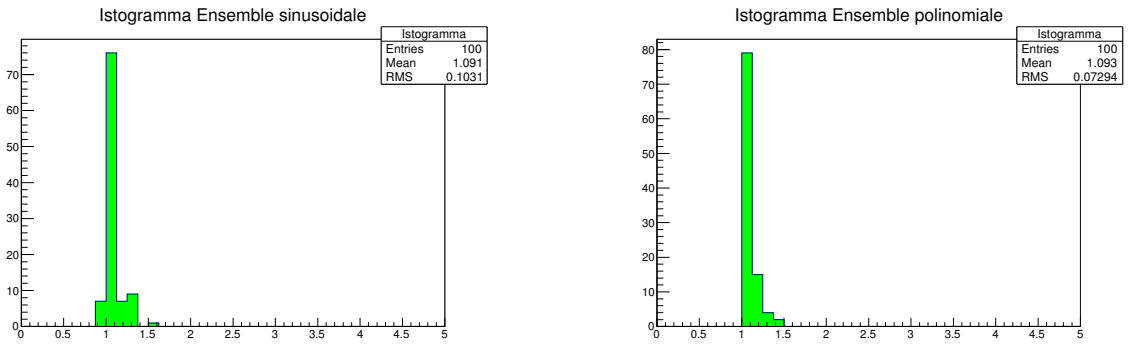

Figura 6: In figura: a sinistra l'istogramma dei $\overline{\chi}^2$ delle PDFs sinusoidali; i valori del $\overline{\chi}^2$ sono intorno ad 1 con una dispersione di circa 10%, indice di una buona performance dell'algoritmo di minimizzazione. A destra l'istogramma dei $\overline{\chi}^2$ delle PDFs polinomiali; anche in questo caso, seppur con qualche layer esterno (PDF con $\overline{\chi}^2 > 1$) l'algoritmo produce un ensemble di funzioni con $\overline{\chi}^2$ in un intorno di 1.

[8]Definiamo il χ^2 normalizzato della PDF f sui dati sperimentali $\mathbf{y}$ come segue:

$$\overline{\chi}^2(\mathbf{y}|f) \doteq \frac{\chi^2(\mathbf{y}|f)}{n}, \tag{44}$$

dove n è il numero di dati sperimentali usati per il calcolo di $\chi^2(\mathbf{y}|f)$.

Le distribuzioni nelle Figure 4 , 5 e 6 mostrano che l'algoritmo di minimizzazione viene iterato un numero sufficientemente alto di volte in modo da produrre un set di distribuzioni partoniche con χ^2 molto vicino a quello teorico.

3.1.3 Risultati numerici

Questa Sezione è dedicata ad una discussione dei risultati computazionali ottenuti tramite il modello descritto nell'introduzione alla Sezione 3.1 . Per mettere in evidenza le differenze previste dalle espressioni dei pesi analizzeremo diverse simulazioni che si riferiscono a differenti settaggi sperimentali. I grafici che proponiamo si basano sul fatto che dato un insieme di valori, il valor medio è uno stimatore del valore vero soggiacente, infatti: supponiamo di aver eseguito un esperimento i cui risultati sono due vettori $(\mathbf{x}, \mathbf{y})$, con $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ ed incertezze σ_i, su ogni y_i. Consideriamo un ensemble di PDFs $\varepsilon = \{f_k\}_{k=1}^{N_{rep}}$ e supponiamo di aver utilizzato una delle espressioni note per il calcolo dei pesi ω_k. Otteniamo N_{rep} coppie del tipo (f_k, ω_k) , $k = 1, \ldots, N_{rep}$. Fissiamo un punto $(x_i, y_i) \in (\mathbf{x}, \mathbf{y})$ con previsione underlying su y_i data da $f_i^0 \doteq f^0(x_i)$ [9]. In x_i avremo N_{rep} diversi valori assunti dalle N_{rep} PDFs dell'ensemble ε, $\{f_1(x_i), \ldots, f_{N_{rep}}(x_i)\}$; poniamo $f_i^j \doteq f_j(x_i)$, con $j = 1, \ldots, N_{rep}$ e $i = 1, \ldots, n$. Calcoliamo il valore della PDF media in x_i come segue

$$\overline{f}_i \doteq \overline{f}(x_i) \doteq \frac{\sum_{j=1}^{N_{rep}} \omega_j f_i^j}{\sum_{j=1}^{N_{rep}} \omega_j}. \tag{45}$$

La PDF media in x_i si ottiene quindi come media pesata, dati certi pesi ω_j, dei valori assunti dalle PDFs che formano l'ensemble ε, valutate ad x_i fissato. Definiamo l'incertezza associata a $\overline{f}(x_i)$ come segue

$$\Delta_i \doteq \sqrt{\frac{\sum_{j=1}^{N_{rep}} \omega_j (f_i^j - \overline{f}_i)^2}{N_{rep} - 1}}. \tag{46}$$

Usiamo questo indicatore per stimare l'incertezza dei valori medi poichè, se i dati sono distribuiti su una gaussiana, la deviazione standard è uno stimatore del livello di confidenza pari al 68%; cioè nel 68% dei casi un dato sperimentale appartiene all'intervallo di 1-σ dato da $\left[\overline{f}_i - \Delta_i, \overline{f}_i + \Delta_i\right]$.
L'insieme dei punti del tipo $(\overline{f}_i, \Delta_i)$, plottati su un grafico in un intervallo $[a, b]$, forma una "banda" che determina gli andamenti della PDF media e della relativa barra d'errore [10] utilizzabili per confrontare le previsioni di diversi Reweighting e del metodo Hessiano.
Per valutare i risultati finali del Reweighting utilizzeremo un test volto ad analizzare gli indicatori di qualità del fit utilizzato. La prima parte del test si sofferma sull'indicatore χ^2, mentre la seconda sulle incertezze delle PDF medie ricavate come illustrato sopra.
Il test per il χ^2 consiste in un confronto fra il χ^2 della funzione sottostante ai dati sperimentali f^0 con quelli delle PDF medie pesate. Il χ^2 di f^0 è definito nel modo usuale per dati senza matrice di covarianza

$$\chi_0^2 \doteq \sum_{k=1}^{n} \frac{[y_k - f_k^0]^2}{\sigma_k^2} \tag{47}$$

dove (x_k, y_k) è il k-esimo dato sperimentale e σ_k l'incertezza su y_k. Il calcolo del χ^2 per le PDFs medie è dato dalla seguente equazione

$$\chi_r^2 \doteq \sum_{k=1}^{n} \frac{[y_k - \overline{f}_k]^2}{\sigma_k^2}. \tag{48}$$

dove l'indice r in χ_r indica un Reweighting.

[9] Nel nostro caso, siccome abbiamo generato i dati sperimentali tramite l'Eq. (41) la previsione sarà $f^0(x_i) = \sin(x_i)$.
[10] Ovviamente, pesi diversi danno diversi punti $(\overline{f}(x_i), \Delta_i)$, quindi l'equazione NNPDF e di Giele-Keller produrranno bande differenti.

Per un corretto controllo delle incertezze calcolate tramite l' Eq. (46) utilizziamo il seguente test: calcoliamo la probabilità che, ad x_i fissato, il valore teorico f_i^0 sia nell'intervallo di 1-σ dato da $\left[f_i^j - \Delta_i, f_i^j + \Delta_i\right]$, cioè $\mathcal{P}_i \doteq \mathcal{P}\left(f_i^0 \in \left[f_i^j - \Delta_i, f_i^j + \Delta_i\right]\right)$. I valori di x_i usati nel calcolo di $\mathcal{P}_i$ sono ottenuti come valori medi fra due punti consecutivi x_j e x_k appartenenti al vettore di dati sperimentali $\mathbf{x}$; così facendo mostriamo non solo che le repliche fittano bene i dati sperimentali, ma anche che interpolano correttamente ogni coppia di valori adiacenti di $\mathbf{x}$. Se il metodo utilizzato fosse corretto, dovrebbe essere che $\mathcal{P}_i = 68\%$ stiamo usando la media pesata come stimatore del valore vero e, supponendo che i valori assunti dalle PDFs a x fissato siano distribuiti su una gaussiana, l'intervallo 1-σ indica che nel 68% dei casi il valore misurato (quindi quello assunto da una PDF) appartiene a tale intervallo. Disporremo quindi di n valori $\{\mathcal{P}_1, \ldots, \mathcal{P}_n\}$, da cui calcoliamo il valor medio $\mathcal{P}$ e la deviazione standard della media che riporteremo nelle tabelle finali sulla discussione dei risultati [11]. Si noti che il test appena esposto può essere utilizzato anche per verificare la bontà di un ensemble ottenuto fittando i dati sperimentali su un dato intervallo prima del Reweighting e di un esnemble ottenuto tramite il metodo Hessiano; in questi casi valori medi ed incertezze diventano valori aritmetici. Il Reweighting che darà un χ^2_m più vicino al valore assunto da χ^2_0 e un valore medio delle $\mathcal{P}_i$ più compatibile con il 68% sarà il modello che meglio riproduce l'andamento delle PDFs medie e che meglio descrive le incertezze della teoria soggiacente ai dati sperimentali.
Di seguito discutiamo nel dettaglio i risultati computazionali. Nei grafici in cui compaiono i dati sperimentali plottiamo quattro PDFs medie: la PDF media-NNPDF (■), la PDF media-Giele-Keller (■), la PDF media-prior (■) e la PDF media-refitted (■). Col termine *PDF media-prior* ci riferiamo alla PDF media calcolata come media delle PDFs dell'ensemble ε ottenute fittando i dati prima dell'implementazione del Reweighting, mentre con *PDF media-refit* alla PDF calcolata mediando le PDFs nell'ensemble ottenuto fittando i dati usati per il prior uniti a quelli usati per il Reweighting [12].

[11] Le incertezze degli indicatori di probabilità verranno usate per evidenziare la compatibilità dei valori di $\mathcal{P}$ nei casi in cui vi siano oscillazioni intorno al 68%.
[12] Si noti che in questi due casi l'Eq. (45) diviene una media aritmetica delle $f_k(x_i)$.

- **Presentazione dei risultati numerici**

L'esempio che portiamo si riferisce ai seguenti settaggi sperimentali, in tabella: "N^o primo dataset" si riferisce al numero di dati sperimentali usati per fittare l'ensemble di PDFs, "1^0 Intervallo" è l'intervallo in cui abbiamo generato i dati per fittare le PDFs, "N^o nuovo dataset " è il numero di dati sperimentali usati per implementare il Reweighting, "2^0 Intervallo" è l'intervallo in cui sono distribuiti i nuovi dati sperimentali, "σ dati" è l'incertezza dei dati sperimentali e "Grado polinomio" è il grado dei polinomi usati nell'esempio numerico.

Descrizione	Valore
N^o primo dataset	100
1^0 Intervallo	[0,3π/2]
N^o nuovo dataset	30
2^0 Intervallo	[3π/2,2π]
σ dati	0.5
Grado polinomio	4

Ensemble sinusoidale

Prima di illustrare i risultati del Reweighitng, è necessario accertarsi del fatto che l'ensemble di PDFs fittato sull'intervallo $[0, 3\pi/2]$ sia costituito da PDFs i cui valori del χ^2 non si discostano molto da quello teorico; questo è stato in parte già mostrato qualitativamente nella Sezione 3.1.2 in cui abbiamo discusso delle performance dell'algoritmo di minimizzazione. Più nel dettaglio, usiamo il test statistico sopra discusso sia sull'ensemble prior (PDFs fittate solo nel 1^0 intervallo) che sul refit (PDFs fittate sull'unione dei dati del 1^0 e del 2^0 intervallo), in modo da confrontare i miglioramenti del secondo ensemble rispetto al primo. Riportiamo i valori dei χ^2 calcolati nell'intervallo $[0, 3\pi/2]$ usato per fittare le PDFs prima del Reweighting; inoltre, altri indicatori della bontà dei fit come le incertezze medie [13] del prior e del refit ed i valori degli indicatori di probabilità $\mathcal{P}$ discussi nell'introduzione di questa Sezione. In Figura 7 sono riportati la PDF media-prior e refit dell'ensemble sinusoidale.

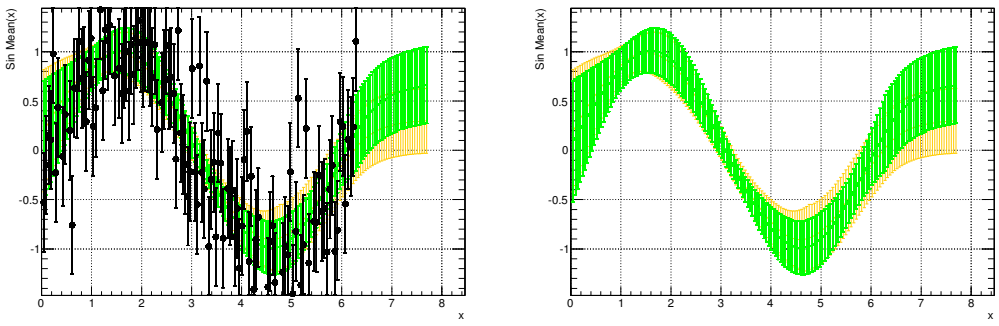

Figura 7: Gli andamenti delle PDFs medie prior e refit dell'ensemble sinusoidale con i dati sperimentali (sinistra) e senza (destra) con 1000 iterazioni: Si nota che nell'intervallo in cui il prior non fitta dati sperimentali, $[3\pi/2, 2\pi]$, la banda gialla presenta delle incertezze leggermente maggiori rispetto a quella verde; un comportamento medesimo si osserva nell'intervallo $[2\pi, 5\pi/2]$ per entrambi gli ensembles.

[13] Con incertezza media intendiamo la media delle Δ_i calcolate nell'Eq. (46).

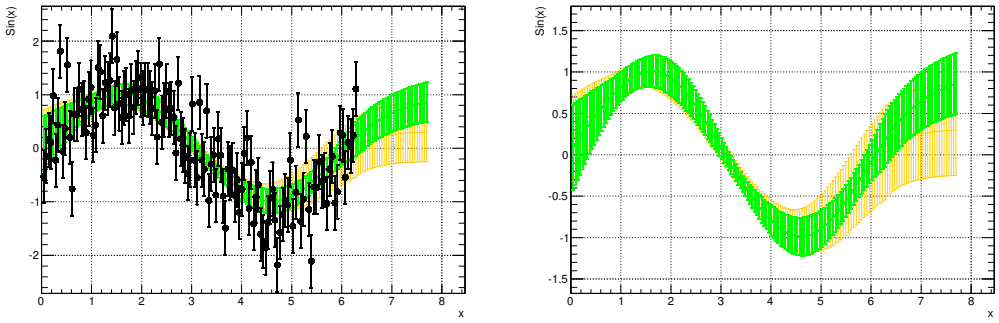

Figura 8: Gli andamenti delle PDFs medie prior e refit dell'ensemble sinusoidale con i dati sperimentali (sinistra) e senza (destra) con 3000 iterazioni: Rispetto a 1000 iterazioni si osserva una riduzione delle incertezze del prior e del suo andamento medio, come verrà mostrato nelle tabelle seguenti.

Come discusso nella Sezione 3.1, implementiamo il test statistico per verificare la qualità dei fit usati con diversi del numero di iterazioni dell'algoritmo genetico. Per il prior ed il refit dell'ensemble sinusoidale otteniamo i seguenti risultati [14] :

Descrizione	$N_{it} = 300$	$N_{it} = 600$	$N_{it} = 1000$	$N_{it} = 2000$	$N_{it} = 3000$
Intervallo	$[0, 3\pi/2]$	$[0, 3\pi/2]$	$[0, 3\pi/2]$	$[0, 3\pi/2]$	$[0, 3\pi/2]$
χ^2_0	101.02	101.02	101.02	101.02	101.02
χ^2_{prior}	112.51	105.31	102.63	101.12	100.81
χ^2_{refit}	109.32	101.48	101.24	100.71	100.62
Δ_{prior}	1.02	0.76	0.41	0.42	0.42
Δ_{refit}	0.89	0.67	0.42	0.39	0.40
$\mathcal{P}_{prior}$	$(78 \pm 4)\%$	$(76 \pm 2)\%$	$(68 \pm 1)\%$	$(69 \pm 1)\%$	$(68 \pm 1)\%$
$\mathcal{P}_{refit}$	$(73 \pm 3)\%$	$(72 \pm 2)\%$	$(68 \pm 1)\%$	$(67 \pm 1)\%$	$(68 \pm 1)\%$

Le prime due righe della tabella relativa all'intervallo $[0, 3\pi/2]$ hanno gli stessi parametri poiché si riferiscono al fatto che l'intervallo usato è sempre $[0, 3\pi/2]$ ed il χ^2 della funzione soggiacente è sempre quello riportato nella prima cella in quanto il settaggio sperimentale è sempre il medesimo. Si noti che per $N_{it} = 3000$ il χ^2 del prior e del refit sono minori di quello della funzione soggiacente, questo è dovuto ad un leggero overlearning degli ensembles: in particolare, siccome la forma funzionale ha solo 4 parametri da determinare, è possibile che 3000 iterazioni siano sufficienti affinché le PDFs inizino a fittare rumore risentendo troppo delle perturbazioni dei dati. Avendo discusso gli indicatori di qualità del prior sull'intero intervallo in cui questo ensemble è fittato, analizziamo gli indicatori del refit su $[0, 2\pi]$ per avere un confronto sui miglioramenti ottenuti.

Descrizione	$N_{it} = 300$	$N_{it} = 600$	$N_{it} = 1000$	$N_{it} = 2000$	$N_{it} = 3000$
Intervallo	$[0, 2\pi]$	$[0, 2\pi]$	$[0, 2\pi]$	$[0, 2\pi]$	$[0, 2\pi]$
χ^2_0	133.05	133.05	133.05	133.05	133.05
χ^2_{refit}	143.03	133.65	133.11	130.93	130.71
Δ_{refit}	0.83	0.65	0.44	0.41	0.40
$\mathcal{P}_{refit}$	$(73 \pm 3)\%$	$(70 \pm 2)\%$	$(68 \pm 1)\%$	$(67 \pm 1)\%$	$(68 \pm 1)\%$

Si noti che all'aumentare della lunghezza d'iterazione gli indicatori di qualità del refit migliorano e, a N_{it} fissato, i parametri del refit su $[0, 2\pi]$ sono migliori dei rispettivi valori del prior sul suo insieme

[14] Nelle tabelle seguenti, le Δ riportate indicano il doppio delle deviazioni standard calcolate tramite l'Eq. (46); usiamo questa notazione per riferirci all'intera larghezza delle bande riportate nelle figure.

di fit ($[0, 3\pi/2]$). Gli ultimi 3 valori degli indicatori di probabilità, seppur con qualche oscillazione, sono compatibili fra loro. Per assicurarci del fatto che entrambi gli ensembles interpolino bene i dati sperimentali, riportiamo gli stessi indicatori della tabella precedente calcolati nell'intervallo di Reweighting $[3\pi/2, 2\pi]$ e in un intervallo di estrapolazione in cui non sono presenti dati sperimentali come $[2\pi, 5\pi/2]$.

Descrizione	$N_{it} = 300$	$N_{it} = 600$	$N_{it} = 1000$	$N_{it} = 2000$	$N_{it} = 3000$
Intervallo	$[3\pi/2, 2\pi]$	$[3\pi/2, 2\pi]$	$[3\pi/2, 2\pi]$	$[3\pi/2, 2\pi]$	$[3\pi/2, 2\pi]$
χ^2_0	32.03	32.03	32.03	32.03	32.03
χ^2_{prior}	46.64	39.76	36.12	34.27	33.43
χ^2_{refit}	33.71	32.17	31.87	30.22	30.09
Δ_{prior}	1.10	0.81	0.51	0.44	0.44
Δ_{refit}	0.77	0.64	0.46	0.42	0.42
$\mathcal{P}_{prior}$	$(78 \pm 4)\%$	$(77 \pm 4)\%$	$(68 \pm 2)\%$	$(70 \pm 2)\%$	$(69 \pm 1)\%$
$\mathcal{P}_{refit}$	$(72 \pm 2)\%$	$(64 \pm 2)\%$	$(68 \pm 1)\%$	$(68 \pm 1)\%$	$(68 \pm 1)\%$

Non avendo dati sperimentali in $[2\pi, 5\pi/2]$, riportiamo solo i valori delle incertezze e delle probabilità.

Descrizione	$N_{it} = 300$	$N_{it} = 600$	$N_{it} = 1000$	$N_{it} = 2000$	$N_{it} = 3000$
Intervallo	$[2\pi, 5\pi/2]$	$[2\pi, 5\pi/2]$	$[2\pi, 5\pi/2]$	$[2\pi, 5\pi/2]$	$[2\pi, 5\pi/2]$
Δ_{prior}	1.14	1.06	0.70	0.65	0.64
Δ_{refit}	0.92	0.88	0.65	0.58	0.57
$\mathcal{P}_{prior}$	$(79 \pm 4)\%$	$(77 \pm 4)\%$	$(73 \pm 3)\%$	$(72 \pm 2)\%$	$(72 \pm 2)\%$
$\mathcal{P}_{refit}$	$(74 \pm 3)\%$	$(72 \pm 2)\%$	$(70 \pm 1)\%$	$(68 \pm 1)\%$	$(67 \pm 1)\%$

Scorrendo ogni riga delle tabelle mostrate, si osserva che all'aumentare del numero di iterazioni gli indicatori di qualità dei fit tendono a stabilizzarsi, indice del fatto che l'algoritmo di minimizzazione migliora la qualità del fit usato fino ad una certa lunghezza d'iterazione dopo la quale si raggiunge la condizione ottimale per procedere con l'implementazione del Reweighting. Si noti che per valori di $N_{it} = 2000 \div 3000$ si hanno dei valori dei χ^2 dell'ensemble refit minori del χ^0 della funzione soggiacente; anche in questo intervallo, come in $[0, 3\pi/2]$, si ha un caso di leggero overlearning. Notiamo tuttavia che il χ^2 prior continua a migliorare; questo vuol dire che vi sono alcune repliche per cui non solo non c'è underlearning, ma un numero elevato di iterazioni dell'algoritmo di minimizzazione è imprescindibile affinché si possa raggiungere la convergenza degli indicatori. Questo mostra un limite di un algoritmo di minimizzazione a lunghezza fissata. Tuttavia, non basta avere la sola stabilità del χ^2 per poter procedere col Reweighting, in quanto è possibile che, per un certo valore di N_{it}, un indicatore sia stabile mentre un altro no; da cui la necessità di verificare diversi valori di N_{it} per assicurarsi che tutti gli indicatori abbiamo raggiunto la stabilità avendo cura di evitare l'overlearning delle funzioni; si parla di overlearning [15]; è quindi necessario stimare la lunghezza d'iterazione giusta affinche i fit utilizzati raggiungano la convergenza degli indicatori di qualità. Per gli ensemble prior e refit della forma funzionale sinusoidale si nota, guardando l'andamento degli indicatori riportati nelle tabelle relative agli intervalli d'interesse, che la convergenza si si raggiunge nell'intervallo $N_{it} \in [2000, 3000]$ in entrambi i casi.
A fini illustrativi riporteremo i grafici delle PDFs medie di entrambi gli ensembles per N_{it} e $3N_{it}$ (1000 e 3000 per l'ensemble sinusoidale, 2000 e 6000 per quello polinomiale) in modo da evidenziare le differenze che si possono ottenere nel caso in cui non venga raggiunta la convergenza di tutti gli indicatori.
Si noti che, come illustrato in Figura 7, in entrambi gli ensembles si ha un aumento delle incertezze medie negli intervalli in cui le PDFs non fittano i dati sperimentali, a parità di numero di iterazioni

[15] Si raggiunge lo stato di overlearning quando le PDFs, oltre fittare i dati sperimentali, fittano rumore causando dei valori del χ^2 troppo piccoli.

utilizzato. Ad esempio, in $[3\pi/2, 2\pi]$ con $N_{it} = 1000$(in $[2\pi, 5\pi/2]$ anche per l'ensemble refit), l'incertezza Δ_{prior} è maggiore di Δ_{prior} nell'intervallo $[0, 3\pi/2]$; questo è dovuto al fatto che le PDF del prior sono fittate solo in $[0, 3\pi/2]$, quindi l'algoritmo di minimizzazione non tiene conto di eventuali dati in altri intervalli. Perciò, in $[3\pi/2, 2\pi]$, le PDFs del prior hanno andamenti che determinano una dispersione maggiore in modo da produrre un'incertezza media maggiore; questo si evince soprattutto con i valori di N_{it} per i quali non si è ancora raggiunta la convergenza delle incertezze [16] Δ. La non convergenza delle Δ spiega anche il fatto che gli indicatori di probabilità aumentano all'aumentare delle incertezze, indice del fatto che l'intervallo $\left[f_i^j - \Delta_i, f_i^j + \Delta_i\right]$, in cui andiamo ad effettuare il test per le incertezze, ha una misura sempre maggiore.
I grafici seguenti mostrano i risultati del Reweighting e del metodo Hessiano:

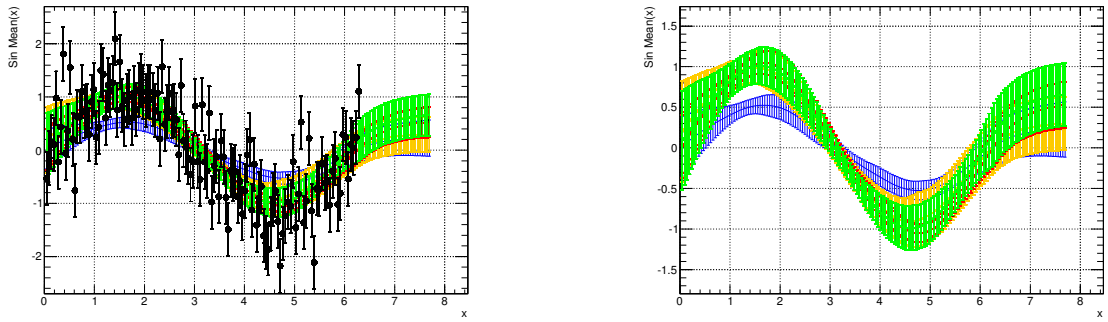

Figura 9: Gli andamenti delle PDFs medie dell'ensemble sinusoidale con i dati sperimentali (sinistra) e senza (destra) con $N_{it} = 1000$: Refit, Prior, NNPDF e Giele-Keller. Qualitativamente si nota che la banda blu si discosta maggiormente da quella rossa e verde che invece presentano andamenti molto simili.

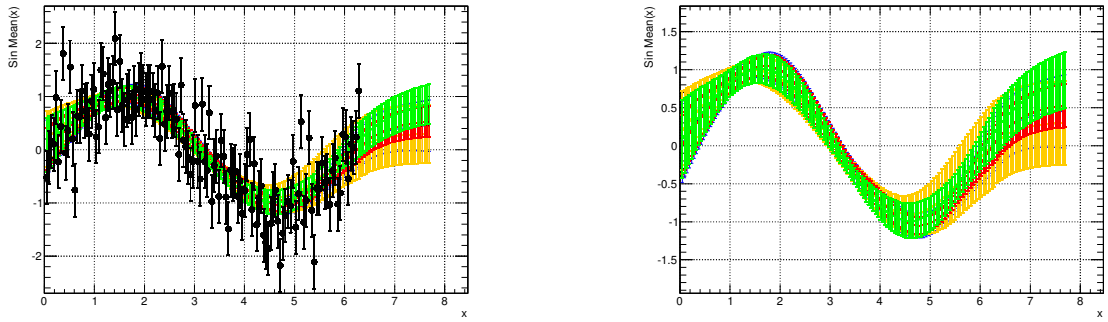

Figura 10: Gli andamenti delle PDFs medie dell'ensemble sinusoidale con i dati sperimentali (sinistra) e senza (destra) con $N_{it} = 3000$: Refit, Prior, NNPDF e Giele-Keller. Si nota un netto miglioramento del valor medio della banda blu e della sua banda d'incertezza come illustrato nelle tabelle precedenti.

Si noti la differenza tra la banda NNPDF ottenuta con 1000 iterazioni (Figura 9) e quella relativa a 3000 (Figura 10). Di seguito discutiamo quantitativamente i risultati del Reweighting tramite il test sul χ^2 e sugli indicatori di probabilità.

[16] Si noti che le incertezze medie delle PDFs, in un caso ottimale, dovrebbero essere minori di quelle sperimentali poiché il fit combina più dati.

Descrizione	$N_{it} = 300$	$N_{it} = 600$	$N_{it} = 1000$	$N_{it} = 2000$	$N_{it} = 3000$
Intervallo	$[0, 3\pi/2]$	$[0, 3\pi/2]$	$[0, 3\pi/2]$	$[0, 3\pi/2]$	$[0, 3\pi/2]$
χ^2_0	101.02	101.02	101.02	101.02	101.02
χ^2_{NNPDF}	124.04	118.41	107.73	104.84	104.79
χ^2_{GK}	103.45	103.02	102.57	101.68	101.54
Δ_{NNPDF}	0.91	0.80	0.45	0.45	0.44
Δ_{GK}	0.78	0.69	0.43	0.40	0.39
$\mathcal{P}_{NNPDF}$	$(52 \pm 3)\%$	$(57 \pm 2)\%$	$(62 \pm 2)\%$	$(67 \pm 1)\%$	$(67 \pm 1)\%$
$\mathcal{P}_{GK}$	$(71 \pm 2)\%$	$(70 \pm 3)\%$	$(68 \pm 1)\%$	$(69 \pm 1)\%$	$(68 \pm 1)\%$

La tabella precedente spiega il miglioramento delle prestazioni del Reweighting NNPDF con l'ensemble sinusoidale: usando 3000 iterazioni tutti gli indicatori di qualità del fit hanno raggiunto dei valori stabili, quindi si ha la convergenza, che nel caso NNPDF, si ha per $N_{it} \in [2000, 3000]$. In modo analogo il Reweighting Giele-Keller raggiunge la convergenza in un intorno di $N_{it} \in [1000, 2000]$. Nell'intervallo $[3\pi/2, 2\pi]$ otteniamo i seguenti risultati:

Descrizione	$N_{it} = 300$	$N_{it} = 600$	$N_{it} = 1000$	$N_{it} = 2000$	$N_{it} = 3000$
Intervallo	$[3\pi/2, 2\pi]$	$[3\pi/2, 2\pi]$	$[3\pi/2, 2\pi]$	$[3\pi/2, 2\pi]$	$[3\pi/2, 2\pi]$
χ^2_0	32.03	32.03	32.03	32.03	32.03
χ^2_{NNPDF}	51.77	46.37	37.71	33.61	30.41
χ^2_{GK}	34.83	33.76	32.12	30.12	30.21
Δ_{NNPDF}	1.04	0.82	0.46	0.43	0.43
Δ_{GK}	0.80	0.69	0.42	0.42	0.40
$\mathcal{P}_{NNPDF}$	$(76 \pm 3)\%$	$(60 \pm 4)\%$	$(60 \pm 3)\%$	$(67 \pm 1)\%$	$(68 \pm 1)\%$
$\mathcal{P}_{GK}$	$(71 \pm 2)\%$	$(69 \pm 3)\%$	$(67 \pm 1)\%$	$(68 \pm 1)\%$	$(68 \pm 1)\%$

Per valutare il modo in cui i due modelli di Reweighting danno previsioni su regioni in cui non disponiamo di dati sperimentali, valutiamo incertezze medie e probabilità nell'intervallo $[2\pi, 5\pi/2]$:

Descrizione	$N_{it} = 300$	$N_{it} = 600$	$N_{it} = 1000$	$N_{it} = 2000$	$N_{it} = 3000$
Intervallo	$[2\pi, 5\pi/2]$	$[2\pi, 5\pi/2]$	$[2\pi, 5\pi/2]$	$[2\pi, 5\pi/2]$	$[2\pi, 5\pi/2]$
Δ_{NNPDF}	1.09	0.88	0.55	0.50	0.49
Δ_{GK}	0.84	0.71	0.46	0.45	0.45
$\mathcal{P}_{NNPDF}$	$(77 \pm 4)\%$	$(75 \pm 4)\%$	$(60 \pm 3)\%$	$(67 \pm 2)\%$	$(67 \pm 1)\%$
$\mathcal{P}_{GK}$	$(74 \pm 3)\%$	$(63 \pm 2)\%$	$(67 \pm 2)\%$	$(69 \pm 1)\%$	$(69 \pm 1)\%$

Basandoci sugli andamenti degli indicatori di qualità dei fit negli intervalli d'interesse, Concludiamo che usando un ensemble di forme funzionali simili a quella soggiacente e lasciando che entrambi i Reweighting raggiungano la stabilità, i Reweighting NNPDF e Giele-Keller sono equivalenti. Si noti che se avessimo usato $N_{it} = 1000$ per entrambi i Reweighting, avremmo commesso un errore in quanto il χ^2_{NNPDF} è abbastanza vicino a quello teorico, mentre tutti gli altri indicatori di qualità NNPDF no. Questo è dovuto al fatto che il Reweighting Giele-Keller raggiunge la stabilità a N_{it} pari a quello del refit, mentre il Reweighting NNPDF necessita di un numero maggiore di iterazioni.

Ensemble polinomiale

La seguente Figura mostra gli andamenti del prior e del refit nell'intervallo $[0, 5\pi/2]$.

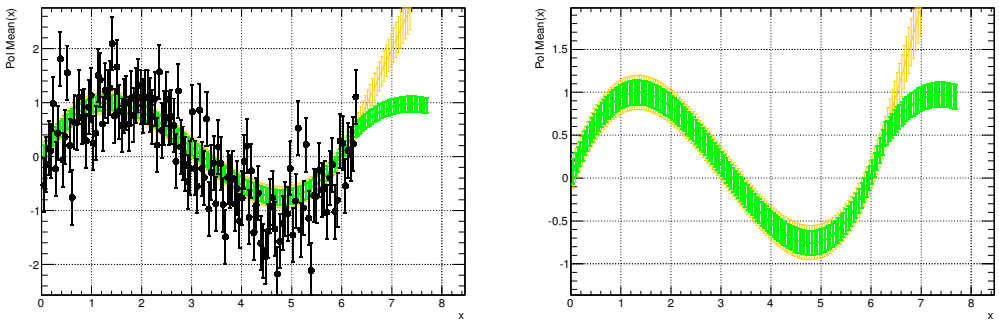

Figura 11: Gli andamenti delle PDFs medie prior e refit dell'ensemble sinusoidale con i dati sperimentali (sinistra) e senza (destra) con 2000 iterazioni: in questo caso, come discusso in Figura 7, la banda gialla presenta delle incertezza molto simili a quelle della banda verde ma, nella regione in cui non vengono fittati dati sperimentali, ha un andamento medio crescente e dà previsioni teoriche peggiori della PDF media-refit.

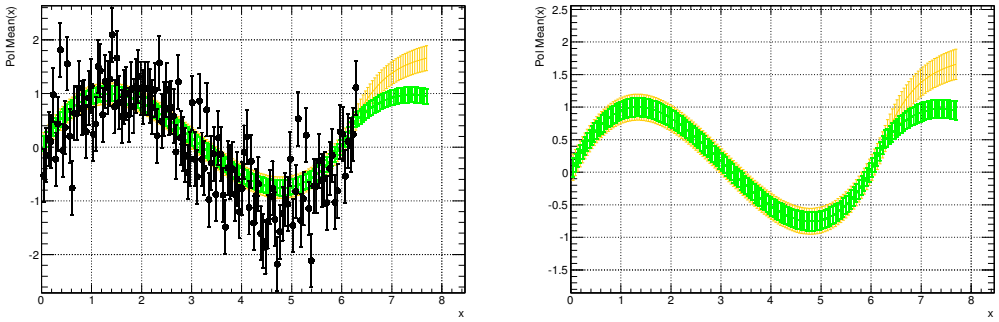

Figura 12: Gli andamenti delle PDFs medie prior e refit dell'ensemble sinusoidale con i dati sperimentali (sinistra) e senza (destra) con 6000 iterazioni: anche per l'ensemble polinomiale si osserva un miglioramento della banda del prior rispetto a 2000 iterazioni.

Usando l'ensemble polinomiale otteniamo i seguenti risultati sul test statistico effettuato prima dell'implementazione del Reweighting e sull'ensemble ottenuto fittando anche i nuovi dati sperimentali:

Descrizione	$N_{it} = 600$	$N_{it} = 1200$	$N_{it} = 2000$	$N_{it} = 4000$	$N_{it} = 6000$
Intervallo	$[0, 3\pi/2]$	$[0, 3\pi/2]$	$[0, 3\pi/2]$	$[0, 3\pi/2]$	$[0, 3\pi/2]$
χ^2_0	101.02	101.02	101.02	101.02	101.02
χ^2_{prior}	121.51	115.36	103.35	101.91	101.12
χ^2_{refit}	112.74	109.18	103.35	100.74	100.51
Δ_{prior}	1.12	0.85	0.42	0.41	0.40
Δ_{refit}	0.94	0.79	0.41	0.39	0.40
$\mathcal{P}_{prior}$	$(77 \pm 3)\%$	$(72 \pm 3)\%$	$(68 \pm 1)\%$	$(66 \pm 2)\%$	$(68 \pm 1)\%$
$\mathcal{P}_{refit}$	$(76 \pm 3)\%$	$(70 \pm 2)\%$	$(68 \pm 1)\%$	$(68 \pm 1)\%$	$(68 \pm 1)\%$

Per l'ensemble refit, nell'intervallo $[0, 2\pi]$ abbiamo i seguenti indicatori di qualità:

Descrizione	$N_{it} = 600$	$N_{it} = 900$	$N_{it} = 1200$	$N_{it} = 4000$	$N_{it} = 6000$
Intervallo	$[0, 2\pi]$	$[0, 2\pi]$	$[0, 2\pi]$	$[0, 2\pi]$	$[0, 2\pi]$
χ^2_0	133.05	133.05	133.05	133.05	133.05
χ^2_{refit}	150.45	143.69	136.81	131.22	130.94
Δ_{refit}	0.92	0.76	0.39	0.41	0.41
$\mathcal{P}_{refit}$	$(75 \pm 4)\%$	$(69 \pm 2)\%$	$(68 \pm 1)\%$	$(67 \pm 1)\%$	$(68 \pm 1)\%$

In tabella si nota una decrescita del χ^2 e degli altri indicatori di qualità in modo da raggiungere la convergenza per $N_{it} \in [4000, 6000]$. Nell'intervallo $[3\pi/2, 2\pi]$ otteniamo:

Descrizione	$N_{it} = 600$	$N_{it} = 1200$	$N_{it} = 2000$	$N_{it} = 4000$	$N_{it} = 6000$
Intervallo	$[3\pi/2, 2\pi]$	$[3\pi/2, 2\pi]$	$[3\pi/2, 2\pi]$	$[3\pi/2, 2\pi]$	$[3\pi/2, 2\pi]$
χ^2_0	32.03	32.03	32.03	32.03	32.03
χ^2_{prior}	40.53	38.69	35.43	34.22	33.38
χ^2_{refit}	37.71	34.51	33.46	30.48	30.43
Δ_{prior}	1.19	0.86	0.40	0.43	0.44
Δ_{refit}	0.91	0.74	0.37	0.42	0.42
$\mathcal{P}_{prior}$	$(78 \pm 5)\%$	$(71 \pm 3)\%$	$(65 \pm 2)\%$	$(65 \pm 3)\%$	$(69 \pm 1)\%$
$\mathcal{P}_{refit}$	$(73 \pm 3)\%$	$(67 \pm 2)\%$	$(69 \pm 1)\%$	$(68 \pm 1)\%$	$(68 \pm 1)\%$

Nell'intervallo $[2\pi, 5\pi/2]$ otteniamo:

Descrizione	$N_{it} = 600$	$N_{it} = 1200$	$N_{it} = 2000$	$N_{it} = 4000$	$N_{it} = 6000$
Intervallo	$[2\pi, 5\pi/2]$	$[2\pi, 5\pi/2]$	$[2\pi, 5\pi/2]$	$[2\pi, 5\pi/2]$	$[2\pi, 5\pi/2]$
Δ_{prior}	1.26	0.90	0.43	0.44	0.43
Δ_{refit}	1.01	0.81	0.41	0.42	0.43
$\mathcal{P}_{prior}$	$(81 \pm 5)\%$	$(76 \pm 3)\%$	$(32 \pm 2)\%$	$(41 \pm 2)\%$	$(42 \pm 1)\%$
$\mathcal{P}_{refit}$	$(75 \pm 4)\%$	$(70 \pm 3)\%$	$(70 \pm 2)\%$	$(69 \pm 1)\%$	$(69 \pm 1)\%$

In questo caso si nota che la convergenza degli indicatori del refit è raggiunta intorno a $N_{it} \in [4000, 6000]$, maggiore del valore dei due Reweighting, ma uguale al valore di convergenza del prior. Si noti che per il prior, nell'intervallo $[2\pi, 5\pi/2]$ in cui non sono presenti dati sperimentali, gli indicatori di probabilità aumentano meno rispetto all'ensemble sinusoidale a N_{it} fissato; questo è dovuto al fatto che, nonostante il debole aumento delle incertezze Δ, stiamo considerando le previsioni di un ensemble polinomiale fittato solo su $[0, 3\pi/2]$; siamo quindi abbastanza distanti dalla regione in cui il prior ha una buona previsione sui dati. Questo motiva l'andamento della PDF media del prior nell'intervallo in cui non sono presenti dati sperimentali e quindi anche i valori bassi degli indicatori di probabilità. Nonostante ciò, usando 6000 iterazioni, negli intervalli in cui sono presenti dati sperimentali si hanno indicatori statistici ottimali per implementare il Reweighting. Analogamente a quanto fatto per l'ensemble sinusoidale, discutiamo i risultati ottenuti con 2000 e 6000 iterazioni.

Di seguito i grafici delle PDFs medie polinomiali:

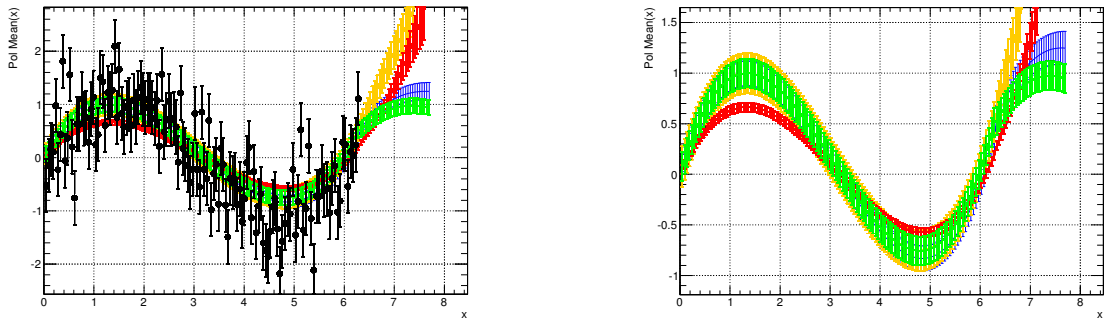

Figura 13: L'andamento delle PDFs medie dell'ensemble polinomiale ottenute con 2000 iterazioni: Refit, Prior, NNPDF e Giele-Keller. Si nota che la banda rossa sottostima in modulo i valori assunti dalle altre e presenta delle incertezze minori su tutto l'intervallo $[0, 5\pi/2]$.

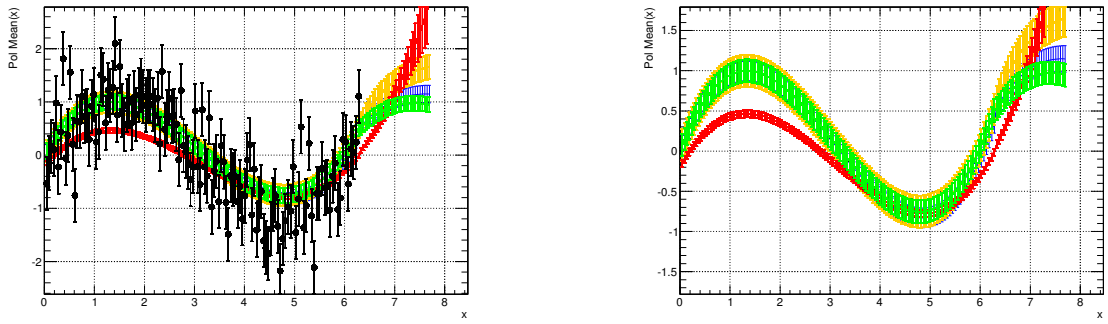

Figura 14: L'andamento delle PDFs medie dell'ensemble polinomiale ottenute con 6000 iterazioni: Refit, Prior, NNPDF e Giele-Keller. Rispetto al caso di Figura 13 si nota un miglioramento del valor medio del prior ed un peggioramento della banda di Giele-Keller.

Di seguito sono riportati i risultati dei test statistici sui modelli di Reweighting:

Descrizione	$N_{it} = 600$	$N_{it} = 1200$	$N_{it} = 2000$	$N_{it} = 4000$	$N_{it} = 6000$
Intervallo	$[0, 3\pi/2]$	$[0, 3\pi/2]$	$[0, 3\pi/2]$	$[0, 3\pi/2]$	$[0, 3\pi/2]$
χ^2_0	101.02	101.02	101.02	101.02	101.02
χ^2_{NNPDF}	109,71	107.48	103.18	101.09	100.89
χ^2_{GK}	117.49	113.16	119.12	123.81	124.14
Δ_{NNPDF}	0.82	0.81	0.41	0.42	0.41
Δ_{GK}	0.41	0.38	0.20	0.20	0.21
$\mathcal{P}_{NNPDF}$	$(72 \pm 3)\%$	$(71 \pm 2)\%$	$(68 \pm 1)\%$	$(68 \pm 1)\%$	$(68 \pm 1)\%$
$\mathcal{P}_{GK}$	$(62 \pm 3)\%$	$(63 \pm 3)\%$	$(51 \pm 2)\%$	$(52 \pm 2)\%$	$(51 \pm 2)\%$

La convergenza del Reweighting NNPDF è raggiunta intorno a $N_{it} \in [4000, 6000]$.

Nell'intervallo $[3\pi/2, 2\pi]$ otteniamo i seguenti risultati

Descrizione	$N_{it} = 600$	$N_{it} = 1200$	$N_{it} = 2000$	$N_{it} = 4000$	$N_{it} = 6000$
Intervallo	$[3\pi/2, 2\pi]$	$[3\pi/2, 2\pi]$	$[3\pi/2, 2\pi]$	$[3\pi/2, 2\pi]$	$[3\pi/2, 2\pi]$
χ^2_0	32.03	32.03	32.03	32.03	32.03
χ^2_{NNPDF}	35.73	33.46	32.61	30.57	30.43
χ^2_{GK}	47.34	45.87	43.37	48.76	49.13
Δ_{NNPDF}	0.85	0.78	0.40	0.41	0.40
Δ_{GK}	0.44	0.41	0.21	0.19	0.21
$\mathcal{P}_{NNPDF}$	$(70 \pm 2)\%$	$(72 \pm 2)\%$	$(67 \pm 1)\%$	$(68 \pm 1)\%$	$(69 \pm 1)\%$
$\mathcal{P}_{GK}$	$(64 \pm 3)\%$	$(61 \pm 3)\%$	$(52 \pm 2)\%$	$(50 \pm 2)\%$	$(50 \pm 2)\%$

Con l'ensemble polinomiale, al contrario di quanto visto per l'ensemble sinusoidale, permettendo agli indicatori di raggiungere la stabilità, il Reweighting Giele-Keller peggiora le previsioni, come indicato dalle tabelle.
Anche in questo caso, i valori $\mathcal{P}_{GK}$ in entrambi gli intervalli (a N_{it} costante) sono dovuti al fatto che la PDF media sottostima il valore soggiacente e presenta una banda di incertezze molto piccola affinché le PDFs possano contenere in un intervallo di 1-σ il valore assunto dalla funzione soggiacente ai dati sperimentali.
Nell'intervallo $[2\pi, 5\pi/2]$ abbiamo che:

Descrizione	$N_{it} = 600$	$N_{it} = 1200$	$N_{it} = 2000$	$N_{it} = 4000$	$N_{it} = 6000$
Intervallo	$[2\pi, 5\pi/2]$	$[2\pi, 5\pi/2]$	$[2\pi, 5\pi/2]$	$[2\pi, 5\pi/2]$	$[2\pi, 5\pi/2]$
Δ_{NNPDF}	0.92	0.80	0.43	0.42	0.42
Δ_{GK}	0.51	0.46	0.31	0.30	0.29
$\mathcal{P}_{NNPDF}$	$(65 \pm 3)\%$	$(76 \pm 4)\%$	$(64 \pm 2)\%$	$(65 \pm 2)\%$	$(66 \pm 2)\%$
$\mathcal{P}_{GK}$	$(66 \pm 3)\%$	$(70 \pm 3)\%$	$(41 \pm 2)\%$	$(41 \pm 1)\%$	$(42 \pm 1)\%$

I valori assunti dagli indicatori statistici nell'intervallo di fitting ed in quello in cui non sono presenti dati sperimentali indicano che usando un ensemble polinomiale il Reweighting NNPDF descrive al meglio le previsioni teoriche ed è in miglior accordo con i risultati prodotti aggiornando l'ensemble tramite il metodo Hessiano.

4 Conclusioni

I test computazionali mostrano i seguenti risultati:

- Il Reweighting NNPDF produce risultati in accordo col modello soggiacente nel caso in cui la forma funzionale scelta per fittare i dati sperimentali sia diversa da quella usata per generarli. Questo si evince dai valori assunti dagli indicatori statistici negli intervalli di interesse: cioè quelli in cui sono presenti i dati sperimentali e quelli subito seguenti che permettono alle PDFs di indicare le previsioni teoriche su regioni non ancora esplorate sperimentalmente.
- Il Reweighting Giele-Keller produce risultati in buon accordo con quelli previsti nel caso in cui i dati sperimentali vengano fittati con una forma funzionale simile a quella soggiacente. Tuttavia, nel caso in cui l'algoritmo di minimizzazione utilizzato sia iterato un numero sufficientemente alto di volte in modo da ottenere indicatori di qualità ottimali per entrambi i Reweighting, pur usando una forma funzionale simile a quella sottostante, il Reweighting NNPDF è statisticamente equivalente al Reweighting Giele-Keller.

Riferimenti bibliografici

[1] S. Forte and G. Watt, "Progress in the Determination of the Partonic Structure of the Proton," Ann. Rev. Nucl. Part. Sci. **63** (2013) 291 [arXiv:1301.6754 [hep-ph]].

[2] S. Forte, "Parton distributions at the dawn of the LHC," Acta Phys. Polon. B **41** (2010) 2859 [arXiv:1011.5247 [hep-ph]].

[3] R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, A. Guffanti, N. P. Hartland and J. I. Latorre *et al.*, "Reweighting and Unweighting of Parton Distributions and the LHC W lepton asymmetry data," Nucl. Phys. B **855** (2012) 608 [arXiv:1108.1758 [hep-ph]].

[4] R. D. Ball *et al.* [NNPDF Collaboration], "Reweighting NNPDFs: the W lepton asymmetry," Nucl. Phys. B **849** (2011) 112 [Erratum-ibid. B **854** (2012) 926] [Erratum-ibid. B **855** (2012) 927] [arXiv:1012.0836 [hep-ph]].

[5] N. Sato, J. F. Owens and H. Prosper, "Bayesian Reweighting for Global Fits," Phys. Rev. D **89** (2014) 114020 [arXiv:1310.1089 [hep-ph]].

[6] H. Paukkunen and P. Zurita, "PDF reweighting in the Hessian matrix approach," arXiv:1402.6623 [hep-ph].

[7] W. T. Giele and S. Keller, "Implications of hadron collider observables on parton distribution function uncertainties," Phys. Rev. D **58** (1998) 094023 [hep-ph/9803393].

5 Ringraziamenti

Rivolgo innanzitutto un ringraziamento al Professor Stefano Forte ed al Dottor Stefano Carrazza per avermi dato l'opportunità di approfondire un argomento di attualità e per la costante disponibilità dimostratami durante tutte le fasi del lavoro.
Rivolgo un ringraziamento di cuore ai miei genitori, Rosa e Girolamo, per il continuo sostegno morale e per i sacrifici che hanno fatto e fanno per permettermi di proseguire gli studi.
Un grazie speciale va a mio fratello Alberto ed alla mia fidanzata Neira per avermi supportato e sopportato in questi tre anni universitari; mi scuso per il tempo sottratto a causa dei mille esami da sostenere.
Infine, un sentito ringraziamento va alla Professoressa Paola Mascherpa per avermi sostenuto nei periodi di difficoltà e per avermi aiutato a crescere.

Fabrizio