PDF DEPENDENCE ON PARAMETER FITS FROM HADRONIC DATA*

Zahari Kassabov

Tif Lab, Dipartimento di Fisica, Università di Milano
and INFN, Sezione di Milano, Via Celoria 16, 20133 Milano, Italy
and
Dipartimento di Fisica, Università di Torino
and INFN, Sezione di Torino, Via P. Giuria 1, 10125, Turin, Italy

(Received January 22, 2018)

We present a discussion on the methods for extracting a given parameter from measurements of hadronic data, with particular focus on determinations of the strong coupling constant. We show that when the PDF dependency on the determination is adequately taken into account, the dispersion between the results from different measurements is significantly reduced. We speculatively propose the concept of *preferred value* of a parameter from a particular dataset.

DOI:10.5506/APhysPolBSupp.11.329

1. Introduction

Since the beginning of its operation, the Large Hadron Collider (LHC) has produced a wealth of experimental data, which have been most notably used to establish the existence of the Higgs boson [1, 2]. These results have not only been used to validate the Standard Model but also to accurately measure its parameters. This involves matching experimental data to theoretical predictions. The experimental results are typically obtained for hadronic cross sections ($\sigma_{pp\to X}$) for a given final state X, while the theory predictions are usually computed for hard (partonic) quantities in the framework of Perturbation Theory, $\hat{\sigma}_{ab\to X}$. The two quantities can be related by universal parton distributions [3]: Using the notation from Ref. [4], we have

$$\sigma_{pp\to X}\left(s, M_X^2\right) = \sum_{a,b} \int_{x_{\min}} \mathrm{d}x_1 \mathrm{d}x_2 f_a\left(x_1, M_X^2\right) f_b\left(x_2, M_X^2\right) \hat{\sigma}_{ab\to X}\left(x_1 x_2 s, M_X^2\right) . \tag{1}$$

* Presented at the Final HiggsTools Meeting, Durham, UK, September 11–15, 2017.

The Parton Distribution Functions (PDFs) of the proton, $f_a(x, Q^2)$, cannot be computed from first principles and instead need to also be determined by appropriately matching experimental data (usually from a wide variety of physical processes) to the predictions for the corresponding partonic cross sections. Thus, a PDF fitting methodology can be viewed as an algorithm that takes as input a set of experimental measurements, together with a set of theory assumptions, and produces a set of parton distribution functions, with an estimate of their uncertainties. Roughly speaking, the PDFs are obtained by minimizing a χ^2 error function

$$\chi^2\left[\{\theta\},\{\alpha\},\mathcal{D}\right] = \sum_{I,J=1}^{N_{\mathcal{D}}} \left(T_I[\{\theta\},\{\alpha\}] - D_I\right) C_{IJ}^{-1} \left(T_J[\{\theta\},\{\alpha\}] - D_J\right) , \quad (2)$$

where $\{\theta\}$ represents the set of parameters that determine the PDF functional form (*e.g.* the neural network parameters in the case of the NNPDF [5] parametrization), $\{\alpha\}$ is the set of theoretical parameters used as input (such as the value of the strong coupling constant $\alpha_{\rm S}(M_Z)$ or the masses of the heavy quarks), $\mathcal{D} = \{D_1 \dots D_{N_{\mathcal{D}}}\}$ is the set of input experimental data points (which will be useful to call *global* dataset), T_I are the theory predictions corresponding to the data $\mathcal{D}_{\mathcal{I}}$, and C_{IJ} is experimental covariance matrix. More realistically, a state of art PDF determination incorporates schemes to compensate [6] biases due to normalization uncertainties [7], regularization mechanisms to avoid overfitting (for example, cross validation [8]) and self validating strategies such as closure tests [8] or tolerances [9, 10]. Usually, the theory parameters $\{\alpha\}$ are kept fixed while the PDF parameters $\{\theta\}$ are optimized. However, it is, in principle, possible to simultaneously optimize for some theory parameter, such as the $|V_{cs}|$ of the CKM matrix [11].

We recall that any theoretical prediction for hadronic observables σ_X, depends trough Eq. (1) both on the choice of PDF fitting methodology and on the data used as input for the PDF fit (even though they are often provided together in the form of a set of PDFs, *e.g.* as grids in the LHAPDF [12] format).

2. Parameter determination from hadronic data

We now turn our attention to the formalism used to extract theory parameters from the best fits to hadronic data. While the discussion applies in general, we restrict ourselves to the determination of the strong coupling constant, $\alpha_{\rm S}$. The value of $\alpha_{\rm S}$ is generally quoted at the mass of the Z boson and is usually [13, 14] taken to be consistent with the *World Average* produced by the Particle Data Group [15]. Two *categories* of determinations

based on hadronic measurements enter the World Average: Those based on PDFs [16–19], which are essentially obtained by optimizing Eq. (2) as a function of α_S, and the $t\bar{t}$ production, currently including only the CMS measurement at 7 TeV [20], which is instead based on minimizing over a χ^2 function that considers explicitly the $t\bar{t}$ data only (we call this the *Partial* χ^2 *method*). We shall discuss the relation between these categories and also try to elucidate the noticeable fact that determinations of α_S based on a hadronic dataset, such as Ref. [20] as well as more recent ones like Ref. [21], give significatively different results from the determinations based on the PDFs that they use as input to compute the predictions in Eq. (1).

2.1. The Partial χ^2 method

Several recent determinations of α_S based on hadronic data [20–25] implement the following procedure, which we shall dub *Partial* χ^2:

1. Consider some experimental measurement of hadronic data, $\mathcal{P}$. For example, $t\bar{t}$ production [20,24], prompt photon events [22] jet production [21,23], and Z+jet production [25].
2. Compute theory predictions at discrete values of α_S, following Eq. (1) and suitably interpolating the results from PDF sets fitted with different values of α_S (*i.e.* where $\alpha_S(M_Z)$ is a fixed parameter in Eq. (2)).
3. Construct a profile $\chi^2_{\mathcal{P}}(\alpha_S)$ characterizing the agreement between data and theory: Analogously to Eq. (2), we have

$$\chi^2_p\left[\alpha_S, \mathcal{P}\right] = \sum_{I,J=1}^{N_{\mathcal{P}}} \left(T_I[\alpha_S] - D_I\right) C^{-1}_{IJ} \left(T_J[\alpha_S] - D_J\right) , \tag{3}$$

   where now the sum is over the partial dataset $\mathcal{P}$.
4. Determine the best fit value of α_S as the minimum of the profile.

We point out that the recommendation [13] for estimating α_S uncertainties on the PDFs, namely obtaining the final result with an upper and a lower PDF variation of $\alpha_S(M_Z)$ does not apply when fitting α_S itself. In this case, the value of α_S should be kept matched with the rest of the calculation. Note that this does not imply that theory parameters cannot be fixed in PDF fits by default: For example, the value of α_S itself is fixed in the PDF4LHC recommendation [13] to a value consistent with the PDG average [15] on the grounds that it takes into account more information than that provided by hadronic data; we may trade some internal consistency of the input $\mathcal{D}$ within the PDF fitting framework with potentially more reliable external constrains on the theory parameters. On the other hand, theoretical parameters that

are to be fitted do certainly have to be varied consistently in the PDFs. This is a required condition, but, as we argue next, not sufficient.

We now discuss the relation between the partial χ^2 method we just described and the dataset used to fit the PDF by optimizing the *global* χ^2, Eq. (2). In particular, it is pertinent to examine why does the partial χ^2 appear to constrain α_S in all the examples above. That is, why is the value of $\chi^2_p[\alpha_S, \mathcal{P}]$ different at different values of α_S?

2.2. PDF and α_S determination from a partial dataset

We note that if the only data used to fit the PDFs was any of the partial datasets above (such as *e.g.* $t\bar{t}$ production), so that $\mathcal{D} = \mathcal{P}$, then we would certainly not have enough constraints to determine the PDFs and α_S simultaneously: In fact, we would be able to obtain an adequate fit, characterized by $\chi^2/(N_\mathcal{D} - 1) \approx 1$ (see Ref. [8] for an extended discussion) for any reasonable value of α_S. We would, however, have big PDF uncertainties, associated to the kinematic regions that are not constrained by $\mathcal{P}$. For example, if we were to fit PDFs to $t\bar{t}$ production data only, we could obtain a good fit at a higher value of $\alpha_S(M_Z)$ by compensating it with a reduced gluon momentum fraction large x as we will show next in a more general situation. Therefore, for $\mathcal{D} = \mathcal{P}$, the partial χ^2 in Eq. (3) is flat and does not allow to determine α_S (in this case, $\chi^2_\mathcal{P}$ is also the global χ^2, Eq. (2)).

It follows that for these relatively small datasets, the $\chi^2_p[\alpha_S, \mathcal{P}]$ profile fundamentally measures the disagreement between the partial data set $\mathcal{P}$ and the dataset included in the PDF fit, $\mathcal{D}$, as a function of α_S.

2.3. Inconsistency of the Partial χ^2 method

The partial χ^2 method neglects the fact that the dataset used in the PDF fits, $\mathcal{D}$, constrains α_S itself, *i.e.* that the minimum of Eq. (2) adopts significantly different values for different values of α_S. That is, given the measurement $\mathcal{P}$, if one makes enough assumptions on the input data of the PDFs to be able to extract $\alpha_S(M_Z)$ with competitive uncertainties, then the prior over α_S is not uniform. One cannot simply disregard the constrains from $\mathcal{D}$ on the theory parameters $\{\alpha\}$ while utilizing them for the PDF parameters $\{\theta\}$. In particular, this can lead to evident inconsistencies such as the value selected by the partial χ^2 method being excluded by the PDF on which the theoretical prediction Eq. (1) is based. This is then a logical contradiction, because the result, which, as we have shown in Sec. 2.2, is based on the agreement with $\mathcal{D}$, is grounded on a prior that is internally inconsistent to begin with. Moreover, the best fit PDFs away from the global minimum in $(\{\alpha\}, \{\theta\})$ are subject to a large degree of arbitrariness: in an ideal PDF fit where all theory and data are correct, every dataset has

a χ^2 per degree of freedom, $\chi^2/\text{d.o.f.} \approx 1$. The χ^2 increases when instead not all the data can be accommodated (*e.g.* because the *wrong* value of α_S has been given as input). In this case, the result of the fit depends on the number of points belonging to each particular dataset, in such a way that the smaller a dataset is (in comparison to others which cannot be fitted simultaneously), the less advantageous it is for the global figure of merit Eq. (2) to bend the PDF in order to accommodate it. This is clear in the case of the $t\bar{t}$ data in the NNPDF 3.1 [5] fits. The default dataset includes a total of 26 $t\bar{t}$ production datapoints corresponding to the ATLAS [26–28] and CMS [29–31] measurements of the total cross sections and differential distributions, computed at NNLO [32, 33] (see Ref. [5] for details). The $t\bar{t}$ data has a large sensitivity to α_S but a low statistical weight in the fit (26 points to be compared to 3979 in total). Therefore, its description (*i.e.* the partial χ^2, Eq. (3)) deteriorates rapidly as we move α_S away from the best fit value. However, we can modify the assumptions on $\mathcal{D}$ insisting that the $t\bar{t}$ data is described at any value of α_S. For example, we set $\alpha_S(M_Z) = 0.121$ where the top data is not so well described in a default NNPDF fit that optimizes Eq. (2) on a large dataset (we have $\chi^2_{t\bar{t}}/\text{d.o.f.} = 1.42$) and increase the statistical weight of the top data by fitting 15 identical copies of it. The effect of the reweighting is to greatly improve the description of $t\bar{t}$ (the partial χ^2 becomes $\chi^2_{t\bar{t}}/\text{d.o.f.} = 1.02$) while slightly deteriorating the global χ^2. The most significative change between the default fit and that with increased weight happens in the gluon PDF, which is nevertheless compatible within PDF uncertainties, as we show in Fig. 1. Indeed, because of the high degeneracy in the space of PDF parameters, $\{\theta\}$, important variations in the input assumptions (that *e.g.* change drastically the partial χ^2) can be reabsorbed into relatively small changes in the PDFs (both in terms of deterioration of the global χ^2 and distances in PDF space). In this way, we have demonstrated that the partial χ^2 does not measure significant physical properties of the hard cross section, but rather properties of the PDF minimization.

In summary, we propose that the most statistically rigorous way to produce an α_S determination from the measurement $\mathcal{P}$ is to include it in a PDF fit and determine simultaneously α_S and the PDFs based on the global χ^2 that now includes $\mathcal{P}$ as well as the rest of the data $\mathcal{D}$. Therefore, if $\mathcal{P}$ was already included in the $\mathcal{D}$, the result from optimizing the global χ^2 profile Eq. (2) would be unchanged. Since there is no way to disentangle the $\mathcal{D}$ dependence from Eq. (1), this method is no more PDF-dependent than the partial χ^2 minimization, but it solves the shortcomings that we have described. The correction on the value of α_S when $\mathcal{P}$ is included will either be a small or point to a flaw in the theory, experiment description or fitting methodology. An important advantage is that the process will then be treated using the full fledged PDF fitting machinery (as opposed to a

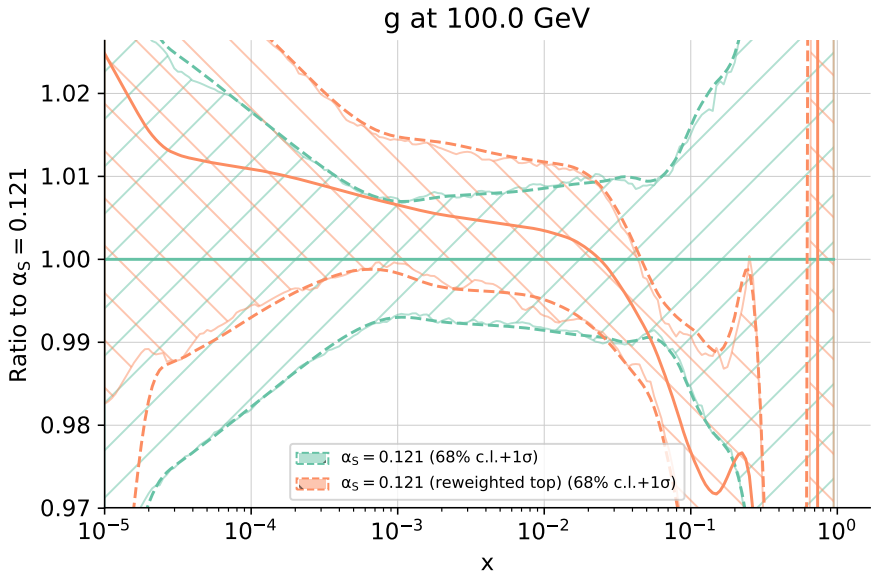

Fig. 1. Comparison of the gluon PDF between an NNPDF 3.1-like global fit at NNLO where we have set $\alpha_S(M_Z) = 0.121$, and another fit where the only difference that the weight with which the $t\bar{t}$ production data enters the fit has been multiplied by 15. The reweighting causes noticeable decrease in the gluon at large x (but yet roughly within uncertainties) to accommodate the $t\bar{t}$ data which is then described optimally, with $\chi^2_{t\bar{t}}/\text{d.o.f.} = 1.02$, to be compared to $\chi^2_{t\bar{t}}/\text{d.o.f.} = 1.42$ before the reweighing. The improvement of the description of the $t\bar{t}$ data comes at the cost of a deterioration in the global χ^2 ($\chi^2/\text{d.o.f.} = 1.215$ before the reweighing and $\chi^2/\text{d.o.f.} = 1.229$ afterwards).

naive minimization of Eq. (2)). In particular, this takes care of implementing the correct treatment of the normalization uncertainties, which has been observed to make a significant difference in an α_S determination [34]. We conclude that it is questionable to consider hadronic results as independent constrains on α_S in World Averages rather than as corrections to the results from the prior PDFs.

3. Preferred values

While we have concluded that the quantities suitable for inclusion in global averages are those based on minimizing the global χ^2 profile, Eq. (2), it is nevertheless interesting to define a *preferred* α_S value from a given dataset $\mathcal{P}$. Some possible usages include the assessment of the constraints provided by the measurement, and possibly the study of the higher order corrections (*e.g.* one could take the dispersion over ensemble of preferred

values of a suitable set of processes as an estimate of Missing Higher Order Uncertainty). We first list some desirable properties that such definition should have.

— Independent of the relation between the number of points in the dataset of interest, $N_{\mathcal{P}}$, and those in the global dataset, $N_{\mathcal{D}}$. Clearly, if we are interested in intrinsic physical properties, the number of points in the dataset should not change the result.

— Explicitly depend on the global dataset used in the PDF fit $\mathcal{D}$. Since, as discussed in the previous section, in general, we cannot get rid of the dependence on $\mathcal{D}$, it needs to be clearly acknowledged.

— Converge to the determination from $\mathcal{P}$ alone, in the sense described in Sec. 2.2, when it determines α_S by itself. While this definition is likely more interesting for smaller, experimentally cleaner datasets, this is a logical asymptotic property.

The Partial χ^2 method discussed in Sec. 2.1 has none of these properties and, therefore, it is not a particularly good definition of preferred value (it may, however, approximate the third property reasonably well in practice). On the other hand, the exercise illustrated in Fig. 1 points at a definition that satisfies them:

Preferred value of α_S for the data P. The value of α_S that corresponds to the minimum of the global χ^2 over values of α_S and PDF parameters $\{\theta\}$, when the PDF parameters are restricted to result in a *good fit* for $\mathcal{P}$ within its experimental uncertainties, for all values of α_S.

The value is preferred in the sense that the constraints from $\mathcal{P}$ take precedence over those from $\mathcal{D}$, in particular, regardless of the number of points, thereby satisfying the first requirement. Once the constrains from $\mathcal{P}$ are enforced, a global χ^2 which includes $\mathcal{D}$ is minimized, thus satisfying the second condition.

The main difficulty is to algorithmically specify what a *good fit* means: Intuitively, if the dataset is self consistent at a given value of α_S, then we require that $\chi^2_{\mathcal{P}}/\text{d.o.f.} \approx 1$. If this is the case at every relevant value of α_S, then the partial $\chi^2_{\mathcal{P}}$ of this reweighed fit is flat and α_S is determined based on the agreement with $\mathcal{D}$ (but based on PDFs that have been modified to accommodate $\mathcal{P}$ at all values of α_S). If $\mathcal{P}$ determines α_S by itself (in the sense of Sec. 2.2), then the partial χ^2 will not be flat and will be used to obtain α_S. A suitable interpolating procedure between these two situations could be obtained in the NNPDF framework by minimizing as a function of $\{\theta\}$ and α_S

$$\text{ERF} = \chi^2\left[\{\theta\}, \alpha_S, \mathcal{D}\right] + w\chi^2\left[\{\theta\}, \alpha_S, \mathcal{P}\right] , \tag{4}$$

where w is a large number. Because of the cross validation-based regularization, the effect of w will saturate either when we reach $\chi^2_{\mathcal{P}}/\text{d.o.f.} \approx 1$, so that only the first therm varies as a function of α_S, or else, if $\mathcal{P}$ determines α_S, the curvature of profile will exclusively depend on the second term.

4. Conclusions

We argue that the uncertainties from determinations of $\alpha_S(M_Z)$ from hadronic data can be significatively reduced by interpreting them as corrections on PDF-based determinations, equivalent to adding the data to the PDF fit. We propose a definition a *preferred value* of a parameter that may be advantageous when studying theory uncertainties.

REFERENCES

[1] G. Aad *et al.* [ATLAS Collaboration], *Phys. Lett. B* **716**, 1 (2012) [arXiv:1207.7214 [hep-ex]].

[2] S. Chatrchyan *et al.* [CMS Collaboration], *Phys. Lett. B* **716**, 30 (2012) [arXiv:1207.7235 [hep-ex]].

[3] R.K. Ellis, W.J. Stirling, B.R. Webber, *QCD and Collider Physics*, Cambridge University Press, 1996.

[4] S. Forte, *Acta Phys. Pol. B* **41**, 2859 (2010) [arXiv:1011.5247 [hep-ph]].

[5] R.D. Ball *et al.* [NNPDF Collaboration], *Eur. Phys. J. C* **77**, 663 (2017) [arXiv:1706.00428 [hep-ph]].

[6] R.D. Ball *et al.* [NNPDF Collaboration], *J. High Energy Phys.* **1005**, 075 (2010) [arXiv:0912.2276 [hep-ph]].

[7] G. D'Agostini, *Nucl. Instrum. Methods Phys. Res. A* **346**, 306 (1994).

[8] R.D. Ball *et al.* [NNPDF Collaboration], *J. High Energy Phys.* **1504**, 040 (2015) [arXiv:1410.8849 [hep-ph]].

[9] S. Dulat *et al.*, *Phys. Rev. D* **93**, 033006 (2016) [arXiv:1506.07443 [hep-ph]].

[10] L.A. Harland-Lang, A.D. Martin, P. Motylinski, R.S. Thorne, *Eur. Phys. J. C* **75**, 204 (2015) [arXiv:1412.3989 [hep-ph]].

[11] R.D. Ball *et al.* [NNPDF Collaboration], *Nucl. Phys. B* **823**, 195 (2009) [arXiv:0906.1958 [hep-ph]].

[12] A. Buckley *et al.*, *Eur. Phys. J. C* **75**, 132 (2015) [arXiv:1412.7420 [hep-ph]].

[13] J. Butterworth *et al.*, *J. Phys. G* **43**, 023001 (2016) [arXiv:1510.03865 [hep-ph]].

[14] D. de Florian *et al.* [LHC Higgs Cross Section Working Group], *CERN Yellow Reports: Monographs*, **2** (2017) DOI:10.23731/CYRM-2017-002 [arXiv:1610.07922 [hep-ph]].

[15] C. Patrignani *et al.* [Particle Data Group], *Chin. Phys. C* **40**, 100001 (2016).

[16] S. Alekhin, J. Blümlein, S. Moch, *Phys. Rev. D* **86**, 054009 (2012).

[17] P. Jimenez-Delgado, E. Reya, *Phys. Rev. D* **79**, 074023 (2009).

[18] L.A. Harland-Lang, A.D. Martin, P. Motylinski, R.S. Thorne, *Eur. Phys. J. C* **75**, 435 (2015).

[19] R.D. Ball *et al.*, *Phys. Lett. B* **707**, 66 (2012).

[20] S. Chatrchyan *et al.* [CMS Collaboration], *Phys. Lett. B* **728**, 496 (2014) [*Corrigendum ibid.* **738**, 526 (2014)] [arXiv:1307.1907 [hep-ex]].

[21] V. Andreev *et al.*, *Eur. Phys. J. C* **77**, 791 (2017) [arXiv:1709.07251 [hep-ex]].

[22] B. Bouzid, F. Iddir, L. Semlala, arXiv:1703.03959 [hep-ph].

[23] M. Aaboud *et al.* [ATLAS Collaboration], *Eur. Phys. J.* **77**, 872 (2017) [arXiv:1707.02562 [hep-ex]].

[24] T. Klijnsma, S. Bethke, G. Dissertori, G.P. Salam, *Eur. Phys. J. C* **77**, 778 (2017) [arXiv:1708.07495 [hep-ph]].

[25] M. Johnson, D. Maître, *Phys. Rev. D* **97**, 054013 (2018) [arXiv:1711.01408 [hep-ph]].

[26] G. Aad *et al.* [ATLAS Collaboration], *Eur. Phys. J. C* **74**, 3109 (2014) [*Addendum ibid.* **76**, 642 (2016)] [arXiv:1406.5375 [hep-ex]].

[27] M. Aaboud *et al.* [ATLAS Collaboration], *Phys. Lett. B* **761**, 136 (2016) [arXiv:1606.02699 [hep-ex]].

[28] G. Aad *et al.* [ATLAS Collaboration], *Eur. Phys. J. C* **76**, 538 (2016) [arXiv:1511.04716 [hep-ex]].

[29] V. Khachatryan *et al.* [CMS Collaboration], *J. High Energy Phys.* **1608**, 029 (2016) [arXiv:1603.02303 [hep-ex]].

[30] V. Khachatryan *et al.* [CMS Collaboration], *Eur. Phys. J. C* **77**, 172 (2017) [arXiv:1611.04040 [hep-ex]].

[31] V. Khachatryan *et al.* [CMS Collaboration], *Eur. Phys. J. C* **75**, 542 (2015) [arXiv:1505.04480 [hep-ex]].

[32] M. Czakon, D. Heymes, A. Mitov, *Phys. Rev. Lett.* **116**, 082003 (2016) [arXiv:1511.00549 [hep-ph]].

[33] M. Czakon, D. Heymes, A. Mitov, *J. High Energy Phys.* **1704**, 071 (2017) [arXiv:1606.03350 [hep-ph]].

[34] NNPDF Collaboration, Precision Determination of the Strong Coupling Constant from a Global PDF Analysis, in preparation.