Eur. Phys. J. C (2024) 84:716
https://doi.org/10.1140/epjc/s10052-024-13100-1

THE EUROPEAN PHYSICAL JOURNAL C

Regular Article - Theoretical Physics

Bayesian inference with Gaussian processes for the determination of parton distribution functions

Alessandro Candido[1], **Luigi Del Debbio**[2], **Tommaso Giani**[3,4,a], **Giacomo Petrillo**[5]

[1] Theoretical Physics Department, CERN, 1211 Geneva 23, Switzerland
[2] Higgs Centre for Theoretical Physics, School of Physics and Astronomy, Peter Guthrie Tait Road, Edinburgh EH9 3 FD, UK
[3] Department of Physics and Astronomy, Vrije Universiteit, 1081 HV Amsterdam, The Netherlands
[4] Nikhef Theory Group, Science Park 105, 1098 XG Amsterdam, The Netherlands
[5] Dipartimento di Statistica, Informatica, Applicazioni "Giuseppe Parenti" (DISIA), Università di Firenze, Viale Morgagni 59, 50134 Firenze, Italy

Received: 7 May 2024 / Accepted: 1 July 2024

Abstract We discuss a Bayesian methodology for the solution of the inverse problem underlying the determination of parton distribution functions (PDFs). In our approach, Gaussian processes (GPs) are used to model the PDF prior, while Bayes' theorem is used in order to determine the posterior distribution of the PDFs given a set of data. We discuss the general formalism, the Bayesian inference at the level of both parameters and hyperparameters, and the simplifications which occur when the observable entering the analysis is linear in the PDF. We benchmark the new methodology in two simple examples for the determination of a single PDF flavor from a set of deep inelastic scattering (DIS) data and from a set of equal-time correlators computed using lattice QCD. We discuss our results, showing how the proposed methodology allows for a well-defined statistical interpretation of the different sources of errors entering the PDF uncertainty, and how results can be validated a posteriori.

1 Introduction

The determination of one or more continuous functions knowing a finite set of experimental observations is notoriously an ill-posed problem, which goes under the name of inverse problem. The extraction of parton distribution functions (PDFs) from experimental and lattice data is an example of this in high-energy physics.

PDFs are an essential input to perform analyses and computations in collider phenomenology, and are required for a number of precision studies concerning the determination of standard model parameters and searches for new physics. PDF determinations currently used for phenomenological studies include MSHT20 [1], CTEQ18 [2], NNPDF4.0 [3], HERAPDF2.0 [4]. A strong dependence on the PDFs has been observed in recent determinations of the strong coupling α_s and of the W boson mass carried out by the ATLAS collaboration [5,6]: when changing the PDF set used as input in the analysis, the output fluctuates by an amount which is bigger than the quoted PDF error. These discrepancies could be generated by some incompatibilities between independent PDF determinations and raises the question whether all the relevant sources of uncertainty are properly accounted for in the quoted PDF error. Comparing the results obtained with different methodologies is one way to test the robustness of the error estimates, and combination studies have been performed to provide the community with a unique PDF set to be used for phenomenology [7].

Given the ill-posed nature of the inverse problem underlying the determination of PDFs, a regularization method is necessary in order to make the problem well defined. The regularization reduces the problem to a finite dimensional and solvable one, but it inevitably introduces some bias, which depends on the specific methodological choices. The solution to an inverse problem will therefore come with an error associated with the methodology, which has to be quantified, just like the other uncertainties entering a PDF fit (uncertainties of the experimental data and of the input standard model parameters, theory errors due to missing QCD higher orders). Despite specific efforts in this direction have already been pursued – for example in Refs. [8,9] the different components of PDF uncertainty are qualitatively assessed using the formalism of closure tests – a way to assess quantitatively the size of the methodological error in a PDF determination is still missing.

[a] e-mail: tgiani@nikhef.nl (corresponding author)

Published online: 22 July 2024

In this paper we investigate a Bayesian approach to the solution of inverse problems, by extending the preliminary work discussed in Ref. [10], and further developing some of the ideas introduced in Ref. [9]. The goal is to develop a methodology, orthogonal to those currently used within the collinear PDF community, where all the relevant sources of uncertainty, including the methodological one, have a clear mathematical definition. We argue that such methodology would simplify the discussion around discrepancies of the kind observed, for instance, in Refs. [5,6], providing a quantitative estimate of the different sources of error entering the PDF uncertainty.

In the fitting methodologies currently used for PDF determinations, the unknown model is parameterized in terms of a finite (albeit large) set of parameters, which are then fitted to the observed data. In the Gaussian Processes approach, rather than starting by a parameterization, a prior probability distribution is introduced for the target model in its original space, encoding our a priori theoretical knowledge of the unknown target function. Using Bayes' theorem, it is possible to determine the posterior distribution of the solution after taking into account a set of experimental observations. This approach has multiple advantages: the inverse problem is well-defined, all the assumptions made on the model are explicitly stated in the choice of the prior and the results are given in terms of posterior probability distributions, making all the relevant uncertainties well-defined from a mathematical point of view. On the other hand, as with any other regularization method, the Bayesian approach introduces a bias through the choice of a specific prior; the posterior probability distribution does depend on the choice of the prior and this dependence needs to be studied and properly quantified. In this paper we will argue that the quantification of the existing bias and the different sources of error affecting the final result is particularly clear in a Bayesian approach.

Our Bayesian approach relies on promoting the values of the PDFs to stochastic variables, whose probability distributions are constrained by experimental data. These posterior probability distributions encode all the information about the PDFs. A possible way to do this is by using the formalism of Gaussian processes (GPs) [11], through which a suitable prior for the unknown PDFs can be defined, in terms of a reduced number of hyperparameters. GPs have already been used to solve inverse problems in various fields in physics, from geophysics [12] to lattice QCD [13–16]. The main focus of this paper is the study of GPs in the context of PDF determinations, including the choice of the most suitable kernel (which defines the prior distribution), the optimization of the corresponding hyperparameters and the way in which theoretical knowledge about PDFs – such as sum rules, kinetic limit, and integrability constraints – can be encoded in the prior.

In Sect. 2 we recall the definition and some well-known properties of GPs, we set the notation and spell out the different steps of the proposed methodology for PDF determination. We focus on the case of observables linear in the PDF and we briefly discuss what changes are required when quadratic observables are included in the analysis. In Sect. 3 we discuss the choice of a prior distribution for PDFs and we provide two simple examples concerning the determination of a single PDF flavor from a set of deep inelastic scattering (DIS) data and lattice equal time correlators. In Sect. 4 we discuss the results, the quantitative evaluation of the different sources of uncertainties entering the analysis, and possible ways to validate the results a posteriori. Conclusions and outlook are presented in Sect. 5.

2 Gaussian processes for inference

In the following, we recall the definition of a Gaussian process, setting the notation for the subsequent sections. Moreover, we describe in detail the case of a GP regression in the presence of data that depend linearly on the GP, subject to a hyperparameterized prior. While the case of linearly dependent data is sufficient for the investigation presented in this work, we also introduce a more general case, which allows us to clarify the simplifications observed in our current study, and sets the framework for further developments towards a global PDF determination. Consequently, we are not going to provide an exhaustive presentation about GPs, for which the reader could refer to the existing literature, such as Ref. [11].

2.1 Notation

In a Bayesian approach, the true value $f(x)$ of the PDF for each $x \in [0,1]$ is treated as a random variable. We should therefore think of x as a continuous index, which parametrizes the elements of a stochastic process. A Gaussian process,

$$f \sim \mathcal{GP}(m,k)\,, \tag{1}$$

is a particular type of stochastic process, whose probability distribution is entirely specified by two functions, the mean $m(x)$ and the kernel $k(x,x')$. The values of the function f at any discrete set of points,

$$\mathbf{x} = \{x_i; i = 1,\dots,N\}\,,$$

define a vector of stochastic variables

$$\mathbf{f} = f(\mathbf{x}) = \begin{pmatrix} f_1 \\ \vdots \\ f_N \end{pmatrix} \in \mathbb{R}^N,\ f_i = f(x_i),\ i = 1,\dots,N\,. \tag{2}$$

The probability distribution of these variables is an N-dimensional Gaussian distribution,

$$\mathbf{f} \sim \mathcal{N}(\mathbf{m}, K)\,, \tag{3}$$

whose mean and covariance are given by

$$\mathbf{m} = m(\mathbf{x})\,, \quad K = k(\mathbf{x}, \mathbf{x}^T)\,, \tag{4}$$

and therefore

$$E[f_i] = m_i = m(x_i)\,, \tag{5}$$

$$\mathrm{Cov}[f_i, f_j] = K_{ij} = k(x_i, x_j)\,. \tag{6}$$

In the following, we will distinguish the points in x for which the value $f(x)$ is included in the theoretical prediction for the measurements, and those where we want to infer the value of the function. We denote the former by $\mathbf{x}$ and the latter by $\mathbf{x}^*$. The corresponding vectors $\mathbf{f}$ and $\mathbf{f}^*$ are defined as in Eq. (2). In a Bayesian formalism, we define a prior joint distribution for $(\mathbf{f}, \mathbf{f}^*)$ and a likelihood function, which will depend on $\mathbf{f}$. We can then compute the posterior distribution for $\mathbf{f}^*$ applying Bayes' theorem. Assuming that

$$\mathbf{f} \in \mathbb{R}^N, \quad \mathbf{f}^* \in \mathbb{R}^M,$$

then the Gaussian process defined in Eq. (1) yields a prior distribution,

$$p(\mathbf{f}, \mathbf{f}^*|\theta) = \frac{1}{\sqrt{\det\,(2\pi\, K)}} \exp\left\{-\frac{1}{2}\left((\mathbf{f}-\mathbf{m})^T, (\mathbf{f}^*-\mathbf{m}^*)^T\right) K^{-1} \begin{pmatrix} \mathbf{f}-\mathbf{m} \\ \mathbf{f}^*-\mathbf{m}^* \end{pmatrix}\right\}, \tag{7}$$

where K is now an $(N+M)\times(N+M)$ matrix,[1]

$$K = \begin{pmatrix} k(\mathbf{x}, \mathbf{x}^T) & k(\mathbf{x}, \mathbf{x}^{*T}) \\ k(\mathbf{x}^*, \mathbf{x}^T) & k(\mathbf{x}^*, \mathbf{x}^{*T}) \end{pmatrix} = \begin{pmatrix} K_{\mathbf{xx}} & K_{\mathbf{xx}^*} \\ K_{\mathbf{x}^*\mathbf{x}} & K_{\mathbf{x}^*\mathbf{x}^*} \end{pmatrix}. \tag{8}$$

The mean and kernel functions might depend on a set of additional parameters, usually referred to as hyperparameters, and collectively denoted as θ. The dependence of the prior on the hyperparameters is marked explicitly in Eq. (7).

2.2 Data and theory predictions

In Ref. [10] we distinguished two different types of input: direct observations of the stochastic process, which we called *point-wise* data, and indirect ones, in which only some functions of the process are actually observed. In this work we will focus on the more general case of indirect observation. In particular, all the results are obtained for a likelihood model in which the data appear as a linear functional of f. Sect. 2.5 describes how to go beyond this assumption.

We denote by T_I the prediction for the I-th datapoint, that will be computed as

$$T_I = \int dx\, c_I(x)\, f(x), \tag{9}$$

where $c_I(x)$ are known functions.[2] The T_I are distributed according to a Gaussian, with mean value and covariance

$$E[T_I] = \int dx\, c_I(x) m(x)\,, \tag{10}$$

$$\mathrm{Cov}[T_I, T_J] = \int dx' dx''\, c_I(x') k(x', x'') c_J(x'') = A_{IJ}. \tag{11}$$

In practice, we are going to be interested in cases where the integral above is computed on a grid of points,

$$T_I = \sum_{i=1}^{N} (\mathrm{FK})_{Ii}\, f_i, \tag{12}$$

where $(\mathrm{FK})_{Ii} = \int c_I(x) p_i(x)$, with $p_i(x)$ an interpolation polynomial, relative to x_i.

The matrix (FK) is called an *FK-table* in the NNPDF jargon, and the notation reflects this convention. Note that the case of point-wise data (direct observation) is obtained in this framework by setting (FK) to the identity. The average and the covariance of the theoretical prediction T induced by the prior probability distribution of $\mathbf{f}$ are given by the discretized versions of Eqs. (10) and (11),

$$E[T_I] = (\mathrm{FK})_{Ij} m_j\,, \tag{13}$$

$$\mathrm{Cov}[T_I, T_J] = (\mathrm{FK})_{Ii}\, (K_{\mathbf{xx}})_{ij}\, (\mathrm{FK})^T_{jJ}. \tag{14}$$

The experimental central value for the data point corresponding to T_I is denoted y_I. Note that T_I is a stochastic variable, while y_I is a constant. In our model, the likelihood is also assumed to be a multivariate Gaussian distribution, and the fluctuations of the data around their central values are described by the *experimental* covariance matrix C_Y.

In the rest of the paper, we will omit the indices like i, j and I, J in the equations above. Boldface vectors, like $\mathbf{f}$ for instance, refer to vectors computed by evaluating the function f on a grid of points. Vectors in the space of data will be denoted by ordinary latin characters; the context should make

[1] We have slightly changed the notation here, compared to the one we used in Ref. [10].

[2] The method exposed in the following works for generic linear functionals, including those that could not be expressed as integrals of regular functions $c_I(x)$. This is the case of the observables analysed in PDF fits, but we limited to this form to simplify the presentation.

it easy to identify these vectors in data space, even though we do not have any typographic convention to identify them.

2.3 Inference for the model

Following the discussion in Ref. [10], we incorporate the knowledge of linear data by introducing the stochastic variable

$$\epsilon \sim \mathcal{N}(0, C_Y), \tag{15}$$

and imposing that

$$(\mathrm{FK})\mathbf{f} + \epsilon = y, \tag{16}$$

where y are the observed experimental central values and C_Y is the covariance matrix of the data. The linear dependence of y on $\mathbf{f}$ is encoded in the matrix (FK).

We are interested in the probability distribution of the vector $\mathbf{f}$ and hyperparameters θ, conditioned on Eq. (16), which we denote as

$$p(\mathbf{f}, \theta|y) = p(\mathbf{f}|\theta, y)\, p(\theta|y). \tag{17}$$

The two factors on the right-hand side of the equation are best analysed separately, since being able to sample both of them is enough to sample the left-hand side. We focus here on the first term, while the second factor will be discussed in the following subsection. The function $p(\mathbf{f}|\theta, y)$ denotes the posterior probability distribution of the vector $\mathbf{f}$ for fixed values of the data and of the hyperparameters θ. In order to compute it, we note that at the level of prior distributions, the vectors $\mathbf{f}$ and $\mathbf{f}^*$, and the data measurement error ϵ must be uncorrelated, hence the covariance matrix describing the joint prior distribution of the three sets of stochastic variables, $(\mathbf{f}, \mathbf{f}^*, \epsilon)$, is a block-diagonal $(N + M + N_{\mathrm{dat}}) \times (N + M + N_{\mathrm{dat}})$ matrix

$$\mathrm{Cov} = \begin{pmatrix} K & 0 \\ 0 & C_Y \end{pmatrix}, \tag{18}$$

where K is the $(N + M) \times (N + M)$ matrix introduced in Eq. (8). Therefore the joint prior is

$$\begin{aligned} p(\mathbf{f}, \mathbf{f}^*, \epsilon|\theta) &= \frac{1}{\sqrt{\det(2\pi K)}} \exp\left\{-\frac{1}{2}\left((\mathbf{f}-\mathbf{m})^T, (\mathbf{f}^*-\mathbf{m}^*)^T\right)\right. \\ &\left. \times K^{-1}\begin{pmatrix}\mathbf{f}-\mathbf{m}\\ \mathbf{f}^*-\mathbf{m}^*\end{pmatrix}\right\} \frac{1}{\sqrt{\det(2\pi C_Y)}} \exp\left\{-\frac{1}{2}\epsilon^T C_Y^{-1}\epsilon\right\}. \end{aligned} \tag{19}$$

Conditioning on the observed values y in Eq. (16),

$$p(\mathbf{f}, \mathbf{f}^*|\theta, y) \propto \int d\epsilon\; p(\mathbf{f}, \mathbf{f}^*, \epsilon|\theta)\, \delta((\mathrm{FK})\mathbf{f} + \epsilon - y) \tag{20}$$

$$\begin{aligned} &\propto \exp\left\{-\frac{1}{2}\left((\mathbf{f}-\mathbf{m})^T, (\mathbf{f}^*-\mathbf{m}^*)^T\right) K^{-1}\begin{pmatrix}\mathbf{f}-\mathbf{m}\\ \mathbf{f}^*-\mathbf{m}^*\end{pmatrix}\right\} \\ &\times \exp\left\{-\frac{1}{2}((\mathrm{FK})\mathbf{f} - y)^T C_Y^{-1}((\mathrm{FK})\mathbf{f} - y)\right\}. \end{aligned} \tag{21}$$

The final step to get $p(\mathbf{f}|\theta, y)$ involves marginalizing $\mathbf{f}^*$, which is readily done remembering that $(\mathbf{f}, \mathbf{f}^*)$ obey a multi-dimensional Gaussian distribution,

$$\begin{aligned} \int d\mathbf{f}^*\, p(\mathbf{f}, \mathbf{f}^*|\theta, y) &\propto \exp\left\{-\frac{1}{2}(\mathbf{f}-\mathbf{m})^T K_{\mathbf{xx}}^{-1}(\mathbf{f}-\mathbf{m})\right\} \\ &\times \exp\left\{-\frac{1}{2}((\mathrm{FK})\mathbf{f} - y)^T C_Y^{-1}((\mathrm{FK})\mathbf{f} - y)\right\} \\ &= \exp\{-S(\mathbf{f}; \theta, y)\}, \end{aligned} \tag{22}$$

so that

$$p(\mathbf{f}|\theta, y) = \frac{\exp\{-S(\mathbf{f}; \theta, y)\}}{\int d\mathbf{f}\, \exp\{-S(\mathbf{f}; \theta, y)\}}. \tag{23}$$

This result was already derived in Eq. (45) in Ref. [9]. Note that

$$\begin{aligned} S(\mathbf{f}; \theta, y) = \frac{1}{2}\Big\{&(\mathbf{f}-\mathbf{m})^T K_{\mathbf{xx}}^{-1}(\mathbf{f}-\mathbf{m}) \\ &+((\mathrm{FK})\mathbf{f} - y)^T C_Y^{-1}((\mathrm{FK})\mathbf{f} - y)\Big\} \end{aligned} \tag{24}$$

is a quadratic form in $\mathbf{f}$, therefore the normalization in Eq. (23) can be computed analytically, yielding a Gaussian posterior for $\mathbf{f}$,

$$p(\mathbf{f}|\theta, y) = \mathcal{N}\left(\mathbf{f}; \tilde{\mathbf{m}}, \tilde{K}_{\mathbf{xx}}\right). \tag{25}$$

Its mean $\tilde{m}$ and covariance $\tilde{K}_{\mathbf{xx}}$ are given by[3]

$$\tilde{\mathbf{m}} = \mathbf{m} + K_{\mathbf{xx}}(\mathrm{FK})^T C_{YT}^{+}\, (\mathbf{y} - (\mathrm{FK})\mathbf{m}), \tag{26}$$

$$\tilde{K}_{\mathbf{xx}} = K_{\mathbf{xx}} - K_{\mathbf{xx}}(\mathrm{FK})^T C_{YT}^{+}\, (\mathrm{FK}) K_{\mathbf{xx}}, \tag{27}$$

where we introduced

$$C_{YT} = (\mathrm{FK}) K_{\mathbf{xx}} (\mathrm{FK})^T + C_Y, \tag{28}$$

which is the covariance of the vector $(\mathrm{FK})\,\mathbf{f} + \epsilon$, and the superscript "+" denotes the matrix pseudoinverse. In the following, we replace the pseudoinverse with the inverse, and the formulae derived implicitly assume that the corresponding matrices are invertible. Equation 27 can be rewritten as

$$\tilde{K}_{\mathbf{xx}}^{-1} = K_{\mathbf{xx}}^{-1} + (\mathrm{FK})^T C_Y^{-1} (\mathrm{FK}). \tag{29}$$

Note that here, and in the rest of this paper, the notation $M_{\mathbf{xx}}^{-1}$ denotes the inverse of the matrix $M_{\mathbf{xx}}$ and not the $(\mathbf{x}, \mathbf{x})$

[3] See [17, ex. 7.4, p. 295] for the proof.

block of the matrix M^{-1}. A more precise notation would be $(M_{\mathbf{xx}})^{-1}$, not to be confused with $(M^{-1})_{\mathbf{xx}}$.

Equations (26) and (29) were already obtained in Ref. [9], while Eq. (27) provides an alternative expression for the posterior covariance, and is the standard formulation in the context of Gaussian processes. Similarly, starting from the same prior and marginalizing with respect to $\mathbf{f}$ we can obtain the posterior for $\mathbf{f}^*$,

$$p\left(\mathbf{f}^*|\theta, y\right) = \mathcal{N}\left(\tilde{\mathbf{m}}*, \tilde{K}^*_{\mathbf{xx}}\right), \tag{30}$$

with

$$\tilde{\mathbf{m}}^* = \mathbf{m}^* + K_{\mathbf{x}^*\mathbf{x}}(\mathrm{FK})^T C_{YT}^{-1}\left(\mathbf{y} - (\mathrm{FK})\mathbf{m}\right), \tag{31}$$

$$\tilde{K}_{\mathbf{x}^*\mathbf{x}^*} = K_{\mathbf{x}^*\mathbf{x}^*} - K_{\mathbf{x}^*\mathbf{x}}(\mathrm{FK})^T C_{YT}^{-1}(\mathrm{FK})K_{\mathbf{xx}^*}. \tag{32}$$

Focusing on the corrections to the mean of the process due to Bayesian inference,

$$\Delta\mathbf{m} = \tilde{\mathbf{m}} - \mathbf{m},$$
$$\Delta\mathbf{m}^* = \tilde{\mathbf{m}}^* - \mathbf{m}^*,$$

we find

$$\Delta\mathbf{m}^* = K_{\mathbf{x}^*\mathbf{x}}K_{\mathbf{xx}}^{-1}\Delta\mathbf{m}, \tag{33}$$

and

$$\tilde{K}_{\mathbf{x}^*\mathbf{x}^*} = K_{\mathbf{x}^*\mathbf{x}^*} - K_{\mathbf{x}^*\mathbf{x}}K_{\mathbf{xx}}^{-1}K_{\mathbf{xx}^*} + K_{\mathbf{x}^*\mathbf{x}}K_{\mathbf{xx}}^{-1}\tilde{K}_{\mathbf{xx}}K_{\mathbf{xx}}^{-1}K_{\mathbf{xx}^*}. \tag{34}$$

Let us emphasise once again that, in this approach, the values of the function f are stochastic variables, and the *information* that we can retrieve about the function at the points $\mathbf{x}^*$ is precisely encoded in the posterior probability distribution. Rather than finding *one* solution, we find the probability distribution of the vector $\mathbf{f}^*$. This is reminescent of what is done when bootstrapping a fit to the data: the posterior distribution in this latter case is the distribution of fit results over the bootstrap sample.

2.4 Inference for the hyperparameters

We now turn to the second term of Eq. (17), namely the posterior of the hyperparameters θ given the data. Using Bayes' theorem we have

$$p(\theta|y) = \frac{p(y|\theta)\, p_\theta(\theta)}{\int d\theta\, p(y|\theta)\, p_\theta(\theta)}, \tag{35}$$

where $p_\theta(\theta)$ denotes the hyperparameters prior. The likelihood $p(y|\theta)$ is proportional to the normalization of the probability distribution $p(\mathbf{f}|\theta, y)$ in Eq. (23), and as such can be computed integrating over $\mathbf{f}$. Alternatively we can get its explicit expression by noticing that the observed data are given by $y = (\mathrm{FK})\,\mathbf{f} + \epsilon$ with

$$(\mathrm{FK})\,\mathbf{f} \sim \mathcal{N}\left((\mathrm{FK})\,\mathbf{m}, (\mathrm{FK})K_{\mathbf{xx}}(\mathrm{FK})^T\right), \tag{36}$$

$$\epsilon \sim \mathcal{N}(0, C_Y), \tag{37}$$

and that therefore

$$y \sim \mathcal{N}((\mathrm{FK})\,\mathbf{m}, C_{YT}), \tag{38}$$

with C_{YT} defined in Eq. (28). In both cases one finds

$$p(y|\theta) = \frac{e^{-\frac{1}{2}(y-(\mathrm{FK})\mathbf{m})^T C_{YT}^{-1}(y-(\mathrm{FK})\mathbf{m})}}{\sqrt{\det\left[2\pi C_{YT}\right]}}. \tag{39}$$

Note that the inference for the model, which yields the posteriors $p(\mathbf{f}|\theta, y)$ and $p(\mathbf{f}^*|\theta, y)$, is completely analytical. This second inference step, which determines the posterior $p(\theta|y)$, in general cannot be solved analytically: because of the hyperparameters appearing in the square root in Eq. (39), the denominator of Eq. (35) cannot be computed analytically any longer. Moreover $p(y|\theta)$, as a function of θ, in general is not a conveniently analyzable probability density function that we know how to sample from. Therefore, this step has to be addressed as a standard inference problem. We can then follow two approaches:

- select the hyperparameter values as the mode of the posterior $p(\theta|y)$:

$$\theta_{\mathrm{MAP}} = \arg\max_\theta p(\theta|y), \tag{40}$$

- use an MCMC algorithm to sample from the posterior $p(\theta|y)$.

Using this second approach the uncertainty due to hyperparameter selection is incorporated into the final PDF uncertainty, and PDFs fitting is reduced to a Monte Carlo problem.

While finding θ_{MAP} is computationally less demanding than MCMC, we opt for the full inference because of our focus on uncertainty quantification. Stopping at a single "best" value can dramatically alter the posterior variance, see, e.g., [18, Fig. 18.18, p. 713].

2.5 The quadratic case

So far we have considered the case in which the theoretical predictions are linear in f_i, according to Eq. (12). In this section we discuss what changes when we consider observables with quadratic dependence on $\mathbf{f}$. Denoting the FK-table for

a quadratic observable T^{quad} as $\widehat{(\text{FK})}$, Eq. (12) becomes

$$T_I^{\text{quad}} = \sum_{i,j=1}^{N} \widehat{(\text{FK})}_{Iij} f_i f_j. \tag{41}$$

The prior given in Eq. (19) remains unchanged, but conditioning now results in

$$\mathbf{f}^T \widehat{(\text{FK})} \mathbf{f} + \epsilon = y, \tag{42}$$

so that, following the same steps as in Sect. 2.3, the posterior $p(\mathbf{f}|\theta, y)$ is now given by

$$p(\mathbf{f}|\theta, y) \propto \exp\left\{-\widehat{S}(\mathbf{f}; \theta, y)\right\}, \tag{43}$$

with

$$\widehat{S}(\mathbf{f}; \theta, y) = \frac{1}{2}\Big\{(\mathbf{f}-\mathbf{m})^T K_{\mathbf{xx}}^{-1}(\mathbf{f}-\mathbf{m}) + (\mathbf{f}^T \widehat{(\text{FK})}\mathbf{f} - y)^T C_Y^{-1}(\mathbf{f}^T \widehat{(\text{FK})}\mathbf{f} - y)\Big\}. \tag{44}$$

The full posterior $p(\mathbf{f}, \theta|y)$ can be written using Bayes' theorem as

$$p(\mathbf{f}, \theta|y) = \frac{\exp\left\{-\widehat{S}(\mathbf{f}; \theta, y)\right\} p_\theta(\theta)}{\int d\theta d\mathbf{f} \exp\left\{-\widehat{S}(\mathbf{f}; \theta, y)\right\} p_\theta(\theta)}. \tag{45}$$

By multiplying numerator and denominator by the marginal likelihood

$$p(y|\theta) \propto \int d\mathbf{f} \exp\left\{-\widehat{S}(\mathbf{f}; \theta, y)\right\} \tag{46}$$

we can recast Eq. (45) in the same form as Eq. (17) with

$$p(\mathbf{f}|\theta, y) = \frac{\exp\left\{-\widehat{S}(\mathbf{f}; \theta, y)\right\}}{p(y|\theta)} \quad \text{and} \quad p(\theta|y) = \frac{p(y|\theta)\, p_\theta(\theta)}{\int d\theta p(y|\theta)\, p_\theta(\theta)}. \tag{47}$$

The difference with respect to the linear case is that Eq. (44) is not quadratic in $\mathbf{f}$. It follows that the posterior $p(\mathbf{f}|\theta, y)$ is not a Gaussian any longer and the likelihood $p(y|\theta)$ cannot be computed analytically.

Both the inference on the parameters $\mathbf{f}$ and on the hyperparameters θ has therefore to be performed at the same time by running a MCMC algorithm starting from Eq. (45). No simplifications occur, and the dimension of the Monte Carlo to run, corresponding to $\dim(\theta)$ in the linear case, is now $\dim(\mathbf{f}) + \dim(\theta)$. This very same approach would work for a generic functional of f.

3 Gaussian processes for PDFs

In the following we apply the formalism described in the previous section to two concrete examples. We consider only the simpler case of observables linear in $\mathbf{f}$, where the inference on the parameters can be done analytically. We first show how, by a suitable choice of the prior, we can implement known physical constraints such as the kinematic limit, sum rules and small-x behaviour. We then give a complete example of the workflow, by determining the nonsinglet triplet PDF

$$T_3 = u^+ - d^+,$$

using a set of synthetic data, first for DIS structure functions, and then for lattice equal-time correlators.

3.1 Prior distribution for T_3

When fitting PDFs we can work in the so-called evolution basis with six non-singlet quark distributions,

$$T_{\mathsf{a}}(x),\ \mathsf{a} = 3, 8, 15, \quad V_{\mathsf{a}}(x),\ \mathsf{a} = 3, 8, 15, \tag{48}$$

the quark singlet distribution, $\Sigma(x)$, and the gluon distribution, $g(x)$. In the following we will be interested in the nonsinglet triplet distribution T_3, to which we associate a GP with zero mean and kernel k

$$T_3 \sim \mathcal{GP}(0, k). \tag{49}$$

The choice of the GPs that define the prior distributions needs to reflect the knowledge of any physical property of the system. We use here a Gibbs kernel [11, p. 93],

$$k(x, y) = \sigma^2 \sqrt{\frac{2l(x)\, l(y)}{l^2(x) + l^2(y)}} \exp\left[-\frac{(x-y)^2}{l^2(x) + l^2(y)}\right] \tag{50}$$

with

$$l(x) = l_0 \times (x + \delta). \tag{51}$$

The quantities σ and l_0 are hyperparameters of the GP, while δ is a small fixed number which regularizes k when $x, y \to 0$. This choice ensures that when approaching the small-x domain the correlation length decreases linearly in x, reflecting the little knowledge we have in this kinematic region. Note that

$$k(x, x) = \sigma^2, \tag{52}$$

which implies a constant amplitude of the kernel on the full x domain. Since we do know something regarding the power behaviour of the PDF at small-x, it would be convenient to

encode it in the prior. This can be done by introducing an additional hyperparameter α and by rescaling the kernel

$$k(x, y) \mapsto \phi(x)\, k(x, y)\, \phi(y), \tag{53}$$

with

$$\phi(x) = x^{\alpha}, \tag{54}$$

so that for the rescaled kernel

$$k(x, x) \propto x^{2\alpha}. \tag{55}$$

In the case of T_3, the PDF has to be integrable in $x = 0$, which can be imposed by choosing $\alpha \in (-1, 0]$. More properties can be implemented in the prior, such as sum rules and kinematic limit, discussed in Appendix A, B.

It should be kept in mind that the choice of the kernel is crucial, and that, when limited experimental data are available, different kernels lead to different results. For the sake of this paper, which aims at presenting the main ideas of the methodology in simple terms, we limit our study to the case of the Gibbs kernel. However, when moving to more complex studies which aim to be used for phenomenology, different choices should be explored, by testing different kernels or defining new kernels suitable for the specific problem of PDF determination. Here follow two ways a particular choice of kernel could turn out wrong in our case. First, the kernel almost completely determines the extrapolation behavior: in this case a mistaken assumption cannot be corrected by the data. Second, and more subtly, it is easy to inadvertently define an ill-conditioned kernel; in other words, a kernel which for practical matters behaves as a finitely parametric model, or that is still "fat" in infinite dimensions, but induces non-zero probability only on some overly specific functions. The textbook example of bad kernel is the exponential quadratic $e^{-(x-y)^2}$, widely used for its simplicity, yet encoding a strong prior; the Gibbs kernel is a variant of it, and so we expect it to have similar problems. Sure that (our own) potential future works will coast along with the Gibbs kernel by inertia, we say, Reader: heed our bootless warning.

3.2 Example 1: T_3 from BCDMS data

Considering DIS on a proton target, the NNLO theory prediction for the structure function F_2^p is

$$F_2^p = C_g \otimes g + C_\Sigma \otimes \Sigma + C_{T_3} \otimes T_3 + C_{T_8} \otimes T_8 + C_{T_{15}} \otimes T_{15}, \tag{56}$$

where $g,\ \Sigma,\ T_3,\ T_8,\ T_{15}$ denote PDF flavors in the so-called evolution basis and C_i the corresponding Wilson coefficients. Considering a neutron target instead and assuming isoscalarity, the neutron PDFs are just the same, except for $u, \bar{u}$ and $d, \bar{d}$, which are exchanged. Since $T_3 = u^+ - d^+$ it follows that the neutron structure function F_2^n can be written as

$$F_2^n = C_g \otimes g + C_\Sigma \otimes \Sigma - C_{T_3} \otimes T_3 + C_{T_8} \otimes T_8 + C_{T_{15}} \otimes T_{15}, \tag{57}$$

with the same Wilson coefficients as in the proton case. Considering a deuterium target, for a generic flavor the corresponding nuclear PDF is

$$f_i^d = \frac{1}{2}\left(f_i^p + f_i^n\right). \tag{58}$$

The deuterium structure function is therefore given by averaging the ones for proton and neutron, getting

$$F_2^d = C_g \otimes g + C_\Sigma \otimes \Sigma + C_{T_8} \otimes T_8 + C_{T_{15}} \otimes T_{15}. \tag{59}$$

Hence we can define the observable

$$F_2^p - F_2^d = C_{T_3} \otimes T_3, \tag{60}$$

which is expressed as the convolution of the Wilson coefficient C_{T_3} with just one PDF, *viz.* T_3. The determination of T_3 using $F_2^p - F_2^d$ only involves one flavor and one FK table, making it an ideal testbed for the method.

Data and FK table Rather than considering real experimental data, we will consider pseudo-data constructed from a known underlying law. This will allow us to test how well the methodology is able to reconstruct the input model (see discussion in Sect. 4). Starting from the datasets BCDMSP and BCDMSD presented in Ref. [19], pseudo-data are generated by identifying points for F_2^p and F_2^d having the same values of the kinematic variables, and taking their difference, which yields a total of 333 points. Following the standard procedure in PDF fits based on factorization, we apply kinematic cuts excluding datapoints where power suppressed corrections could be relevant, leaving 248 points in our analysis. For the experimental error C_Y, we consider the full experimental covariance for the observable $F_2^p - F_2^d$,

$$C_Y = \mathrm{Cov}\left[F_2^p, F_2^p\right] + \mathrm{Cov}\left[F_2^d, F_2^d\right] - 2\mathrm{Cov}\left[F_2^p, F_2^d\right], \tag{61}$$

which is computed using the publicly-available experimental information. As underlying law we could use any functional form we like. To consider a realistic scenario we take the central value of the recent PDF release NNPDF4.0. We denote as y_0 the data generated from the underlying law $\mathbf{f}_0$ using the corresponding (FK) table

$$y_0 = (\mathrm{FK})\, \mathbf{f}_0. \tag{62}$$

The corresponding experimental measurements y entering Eqs. (16), (39) are built as

$$y = y_0 + \eta, \quad \text{with} \quad \eta \sim \mathcal{N}(0, C_Y). \tag{63}$$

Hyperparameters inference Hyperparameters inference is the first step of the procedure and is carried on as described in Sect. 2.4. The hyperparameters entering the analysis are α, l_0 and σ. As a prior we choose a flat[4] distribution having support (0, 10) in the case of l_0 and σ, and $(-1, 0)$ in the case of α, in order to ensure the integrability of T_3 at small-x. To produce results we have run an MCMC algorithm using the Python package `PyMC` [20]: the NUTS sampler is run on 4 independent chains, which are merged after thermalization in a unique set of samples. The posteriors for the three hyperparameters are plotted in Fig. 1. Starting from flat priors, the inference based on the available data generates non-trivial posterior distributions. We will comment more extensively on these results later.

Gaussian inference and results Having determined the posterior of the hyperparameters $p(\theta|y, C_Y)$, we can generate an ensemble of hyperparameters by sampling this distribution. For each hyperparameters sample, the posterior of the parameters $p(\mathbf{f}^*|\theta, y, C_Y)$ can be computed analytically, and a Gaussian replica can be drawn from it. This two-step procedure produces exactly a sample from $p(\mathbf{f}^*, \theta|y, C_Y)$. In Fig. 2 we show the final results, obtained by sampling from $p(\mathbf{f}^*, \theta|y, C_Y)$.

3.3 Example 2: T_3 from lattice data

The same kind of inverse problem as the one presented in the previous section is found when reconstructing PDFs from a discrete set of values for lattice equal-time correlators [15,21,22]. Following Ref. [22], we can reconstruct the distribution T_3 from a set of data for reduced Ioffe-time pseudodistributions [23]. Denoting the latter as $\mathcal{M}\left(\nu, z_3^2\right)$, its imaginary part is related to T_3 by the integral relation

$$\mathrm{Im}\left[\mathcal{M}\right] = \int_0^1 dx\, C^{\mathrm{Im}}\left(x\nu, \mu^2 z_3^2\right) T_3\left(x, \mu^2\right). \tag{64}$$

Also in this case we consider pseudo-data: central values are built according to Eq. (64) using the analytical expression and the kinematic values described in Ref. [22] and NNPDF4.0 as input PDF set. This gives a total of 48 points in the $\left(\nu, z_3^2\right)$ plane. The covariance matrix C_Y in this case is given by the uncertainties coming from the actual lattice simulation and we use here the covariance described in Ref. [22]. Also in this case, Eq. (64) is implemented as per Eq. (12) by means of suitable FK tables. We repeat the same steps as in the previous section, starting from the same prior for T_3 and changing only the FK tables and data entering the framework. The posterior for the hyperparameters and the resulting PDF are plotted in Figs. 3, 4.

3.4 Discussion

Looking at the posteriors plotted in Figs. 1, and 3, we notice how, even though we started from flat priors, the inference based on the available data has generated non-trivial posterior distributions, with more or less sharp peaks depending on the specific case we consider: while BCDMS data give fairly peaked distributions, lattice data seem to allow a broader range of hyperparameter values, especially in the case of α and l_0. As pointed out in Sect. 2.4, the corresponding uncertainty is included in the final results: by sampling from the full posterior $p(\mathbf{f}^*, \theta|y, C_Y)$ the PDF error plotted in Figs. 2, and 4, includes the component due to different possible values of the hyperparameters. Overall, the posterior distributions of the hyperparameters plus the Gaussian sampling performed according to the posterior covariance matrix give a broader distribution at the level of the final PDF in the case of the lattice data. Inspecting Fig. 2 a number of qualitative considerations can be done: in the kinematic region sensitive to the data the input PDF $\mathbf{f}_0$ is reconstructed with a small error, while in the small- and large-x extrapolation region, where no experimental information is available, the posterior strongly depends on the prior we chose in the first place. As discussed in Sect. 3.1 we incorporate in the prior a behavior giving larger error which yet is compatible with integrability properties of T_3. Lacking of any small-x experimental information, this is what we find back in the posterior at small values of x (Fig. 2, right panel). Similarly for $x > 0.8$ we enter the large-x extrapolation region, which is reflected by an increase of the error band visible in the left panel of Fig. 2 (see Appendix B). Similar qualitative considerations can be done for the lattice data case, by looking at the plot in Fig. 4.

In the next section we will make the discussion around these results more quantitative, showing how different components entering the final error can be identified and by introducing different metrics to validate the methodology.

4 Uncertainties and validation

In the following we discuss the final uncertainty of the result, and identify different components associated to the experimental and reconstruction error, the latter being associated to the ill-posed nature of inverse problems. We then discuss the extent to which the underlying model used to generate pseudo-data is reconstructed.

[4] Flat priors are not a good default in general. Here, with only three free hyperparameters, the choice of prior does not matter much; but with the additional model complexity we expect to employ for the full PDF analysis, it will be worth making a reasoned choice.

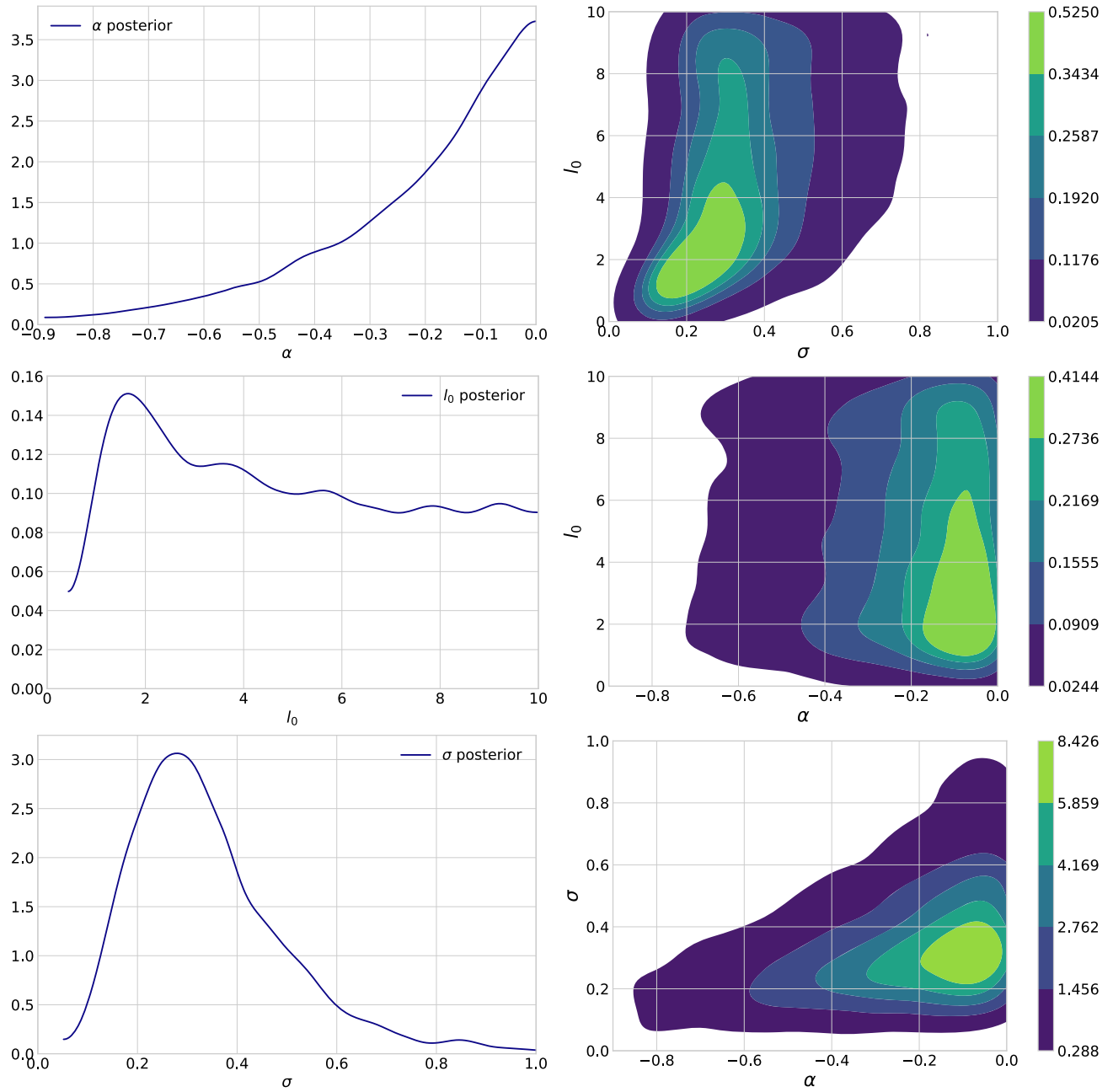

Fig. 1 1-dimensional (left panel) and 2-dimensional (right panel) posteriors of the hyperparameters α, l_0 and σ. The hyperparameter σ is characterized by a sharply peaked posterior located around $\sigma \sim 0.25$ which quickly decays to zero (for this reason in the posterior plot only the region (0, 1) is shown, even if the support of the prior is (0, 10)); α tends to sit closer to 0, with a slow decay for smaller values towards -1; the l_0 posterior discards the smaller values of the correlation length, it shows a peak for $l_0 \sim 1.7$ and then remains fairly constant

4.1 Decomposition of the posterior covariance matrix

Vanishing experimental error Let us first consider the scenario of no experimental error, *i.e.* let us assume that the experimental measurements reproduce the true data with no error. From Eqs. (62), (63) it follows

$$y = y_0 = (\mathrm{FK})\mathbf{f}_0.$$

Following Sec. 3.1.1 of Ref. [12], let us define the resolution kernel as

$$R^{(0)}_{\mathbf{x}^*\mathbf{x}} = K_{\mathbf{x}^*\mathbf{x}}\,(\mathrm{FK})^T \left[(\mathrm{FK})\,K_{\mathbf{xx}}(\mathrm{FK})^T\right]^{-1}(\mathrm{FK})\,. \tag{65}$$

Equation (31) can be rewritten as

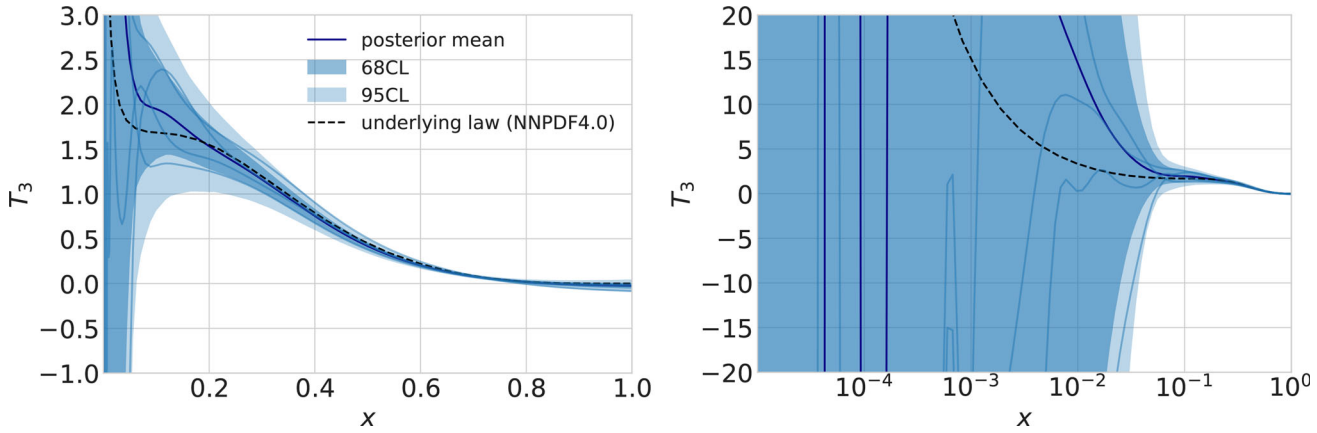

Fig. 2 Samples from $p(\mathbf{f}^*, \alpha, l_0, \sigma | y, C_Y)$ plotted in linear (left panel) and log (right panel) scale. The dark blue line represent the mean of the distribution, while the black dotted line the input PDF NNPDF4.0 used to generate pseudo-data. The shaded regions represent the 68CL and 95CL intervals, and in light blue we plot a few representative samples from the distribution. The posterior displays a smaller variance in the regions sensitive to experimental data, and an increasing spread in the small and large-x extrapolation regions, where the results are mostly determined by the chosen prior

$$\tilde{\mathbf{m}}^* - \mathbf{m}^* = R^{(0)}_{\mathbf{x}^*\mathbf{x}} (\mathbf{f}_0 - \mathbf{m}) \tag{66}$$

which, for $\mathbf{m}^* = \mathbf{m} = 0$, reduces to

$$\tilde{\mathbf{m}}^* = R^{(0)}_{\mathbf{x}^*\mathbf{x}} \mathbf{f}_0. \tag{67}$$

Equation (67) shows that the result of Bayesian inference is a smeared version of the "true" answer $\mathbf{f}_0$, with the smearing kernel given by $R^{(0)}_{\mathbf{x}^*\mathbf{x}}$. The difference between the mean value of the posterior and the underlying law is

$$\tilde{\mathbf{m}}^* - \mathbf{f}_0^* = R^{(0)}_{\mathbf{x}^*\mathbf{x}} \mathbf{f}_0 - \mathbf{f}_0^* \,. \tag{68}$$

We can further specialize the discussion by considering the case $\mathbf{x}^* = \mathbf{x}$ (*i.e.* by looking at the posterior on the x-points of the FK table). In this case we have

$$\tilde{\mathbf{m}} - \mathbf{f}_0 = \left[R^{(0)}_{\mathbf{xx}} - \mathbb{1} \right] \mathbf{f}_0, \tag{69}$$

and the covariance of the posterior can be written as

$$\tilde{K}_{\mathbf{xx}} = \left(\mathbb{1} - R^{(0)}_{\mathbf{xx}} \right) K_{\mathbf{xx}}. \tag{70}$$

Using the FK tables, Eqs. (69), (70) can be recast in terms of bias $\mathcal{B}$ and variance $\mathcal{V}$ in data space, as defined in Ref. [9]. The bias is the difference between the true data y^{true} and the corresponding theory prediction computed using the result of the analysis, and represents the amount by which the resulting model fails in reconstructing the true data. The variance gives the error of the corresponding theory predictions. Writing down their explicit expressions Eqs. (71), (72), it is clear that bias and variance in data space both vanish,

$$\mathcal{B} = (\text{FK}) (\tilde{\mathbf{m}} - \mathbf{f}_0) = (\text{FK}) \left(R^{(0)}_{\mathbf{xx}} - \mathbb{1} \right) \mathbf{f}_0 = 0, \tag{71}$$

$$\mathcal{V} = (\text{FK}) \, \tilde{K} \, (\text{FK})^T = (\text{FK}) \left(\mathbb{1} - R^{(0)}_{\mathbf{xx}} \right) K_{\mathbf{xx}} (\text{FK})^T = 0. \tag{72}$$

In the case of zero experimental error, the methodology reconstructs the input experimental data exactly, independently on the specific values of the hyperparameters; note that despite the fact that there is no bias in data space, the model function is not in general reconstructed exactly, *i.e.* $\tilde{\mathbf{m}} \neq \mathbf{f}_0$ (but they are equal if (FK) has independent columns). Therefore in the case of infinitely precise data, perfect reconstruction is achieved at the data level, but not in the functional space, where a residual reconstruction error is still present.

Non-vanishing experimental error Now let us introduce back a non vanishing experimental error. The reconstruction kernel is

$$R_{\mathbf{x}^*\mathbf{x}} = K_{\mathbf{x}^*\mathbf{x}} (\text{FK})^T \left[(\text{FK}) K_{\mathbf{xx}} (\text{FK})^T + C_Y \right]^{-1} (\text{FK}), \tag{73}$$

and Eqs. (69), (70) become

$$\tilde{\mathbf{m}} - \mathbf{f}_0 = [R_{\mathbf{xx}} - \mathbb{1}] \mathbf{f}_0 + a^T_{\mathbf{xx}} \eta \,, \tag{74}$$

$$\tilde{K}_{\mathbf{xx}} = (\mathbb{1} - R_{\mathbf{xx}}) K_{\mathbf{xx}} (\mathbb{1} - R_{\mathbf{xx}})^T + a^T_{\mathbf{xx}} C_Y a_{\mathbf{xx}} \,, \tag{75}$$

where we have introduced

$$a^T_{\mathbf{xx}} = K_{\mathbf{xx}} (\text{FK})^T \left[(\text{FK}) \, K_{\mathbf{xx}} \, (\text{FK})^T + C_Y \right]^{-1} \,, \tag{76}$$

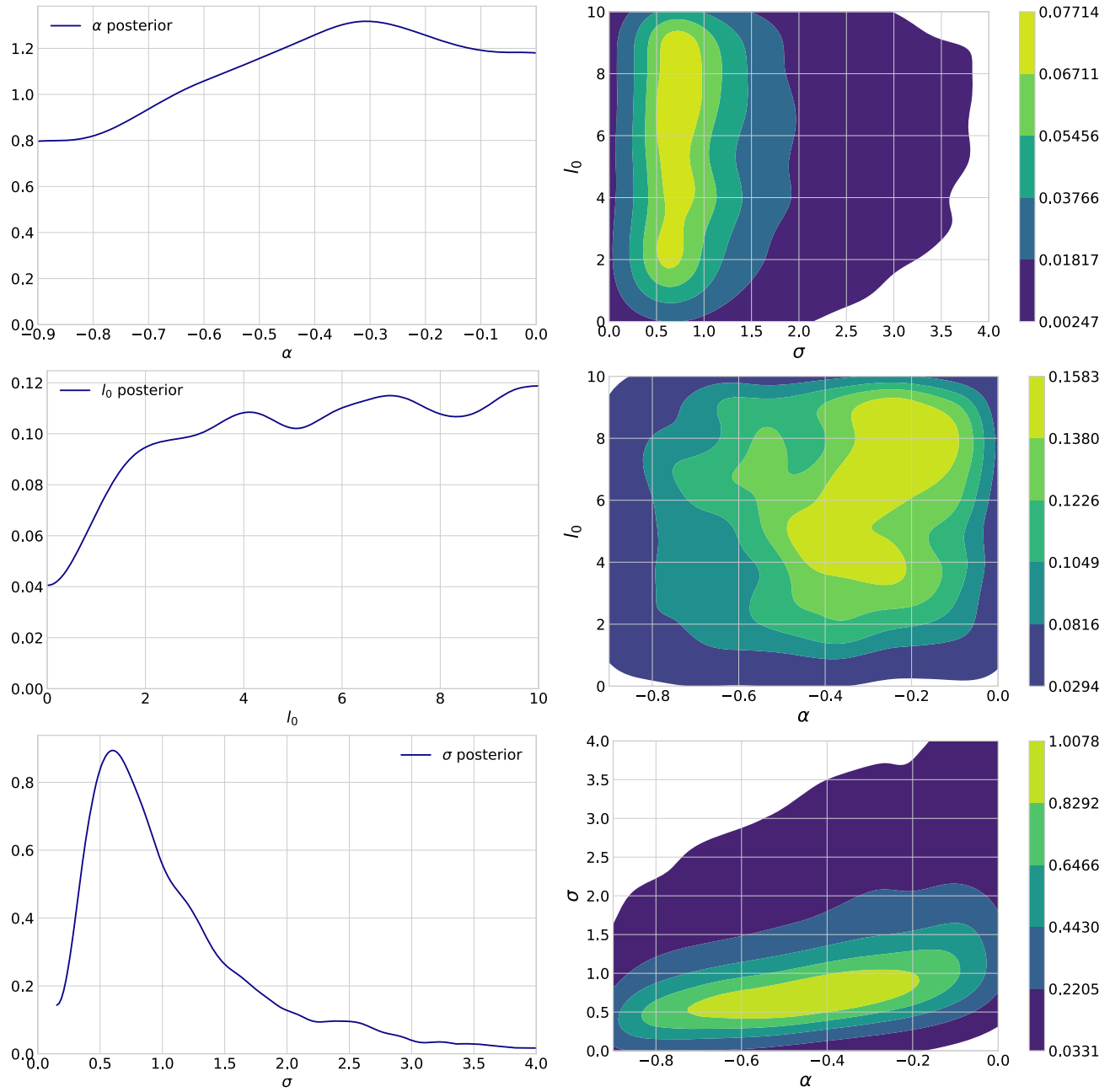

Fig. 3 Same as Fig. 1 for the case of lattice pseudo-data. While the hyperparameter σ is characterized by a sharply peaked posterior distribution, both α and l_0 show fairly flat posteriors, with the one for l_0 only penalizing smaller values of the correlation length

so that

$$R_{\mathbf{xx}} = a_{\mathbf{xx}}^T (\mathrm{FK}). \tag{77}$$

The corresponding expressions in data space Eqs. (71), (72) become

$$\mathcal{B} = (\mathrm{FK})\,[R_{\mathbf{xx}} - \mathbb{1}]\,\mathbf{f}_0 + (\mathrm{FK}) a_{\mathbf{xx}}^T \eta, \tag{78}$$

$$\mathcal{V} = (\mathrm{FK})\,(\mathbb{1} - R_{\mathbf{xx}})\,K_{\mathbf{xx}}\,(\mathbb{1} - R_{\mathbf{xx}})^T\,(\mathrm{FK})^T + (\mathrm{FK}) a_{\mathbf{xx}}^T C_Y a_{\mathbf{xx}} (\mathrm{FK})^T. \tag{79}$$

Perfect reconstruction is not achieved anymore: bias and variance in data space no longer vanish, and their specific value will depend on the choice of the hyperparameters for the kernel. The decomposition in Eqs. (74), (75) highlights the fact that there are two types of contributions to the bias and to the posterior covariance matrix. The first term in Eqs. (74), (75)

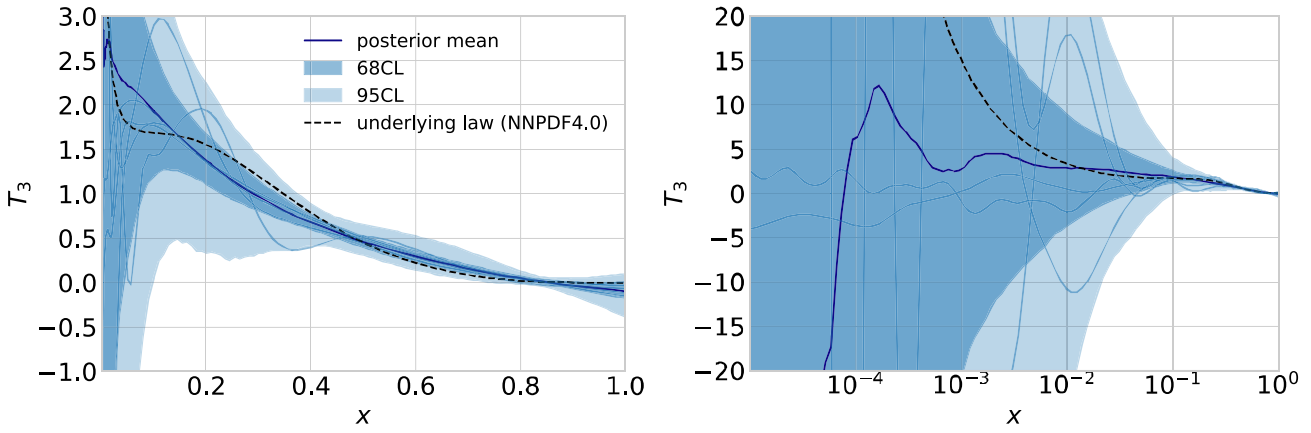

Fig. 4 Same as Fig. 2 for the case of lattice pseudo-data. As in the case of BCDMS data, smaller error are observed in the kinematic regions which are sensitive to the available data. In general, by comparing this plot with Fig. 2, it is clear how lattice data provide less stringent constraints on T_3 than BCDMS. Nevertheless the input PDF is still well reconstructed by the posterior

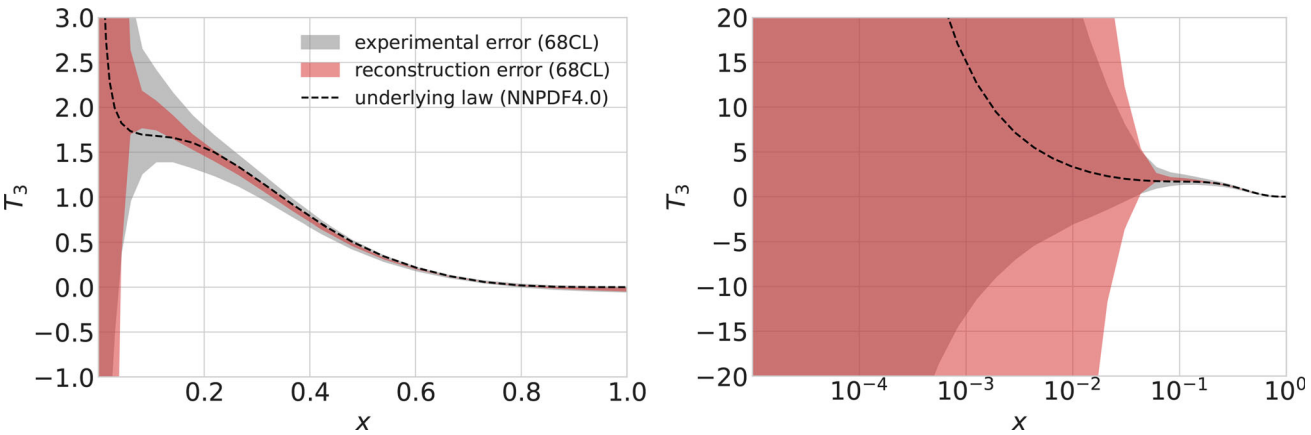

Fig. 5 Same as Fig. 2 with the 68% CL band for the experimental and reconstruction errors plotted separately in grey and red respectively, according to Eq. (75). Note: the error bands are overlapped with transparency, rather than stacked

comes from the limited reconstruction of the central value and indeed would vanish when $R_{\mathbf{xx}} = \mathbb{1}$ [12]. This term survives in the limit where $C_Y \to 0$, *i.e.* in the limit of no experimental errors on the data, in which case we have $R_{\mathbf{xx}} \to R_{\mathbf{xx}}^{(0)}$ and we recover Eqs. (69), (70). The second term is the propagation of the covariance of the data into the covariance of the model. In the case $R_{\mathbf{xx}} = \mathbb{1}$, the only error fluctuations in the posterior distribution come from this term. In Fig. 5, Montecarlo samples generated according to the reconstruction and experimental components are plotted separately for the BCDMS results, in red and grey respectively. In the medium-x region, where more experimental data are available, the PDF uncertainty is dominated by the experimental error, yet a smaller reconstruction error is still present; when moving to the small and large-x extrapolation regions the reconstruction error becomes the dominant one, pointing out the lack of experimental information. We stress once more how these qualitative considerations are precisely quantified in Eqs. (74), (75), giving the analytical expression for the posterior covariance matrices associated to the experimental and reconstruction error, making it possible to quote different component of the PDF error in the context of a pheno analysis.

4.2 Validation

As discussed in the previous section, whenever we deal with noisy experimental information perfect reconstruction of the true data is not achieved anymore and different methodologies are expected to perform differently. In this section we introduce some statistical metrics which allow to validate a given methodology. We discuss the definition of such metrics and what we expect in case of successful reconstruction of the underlying law. Finally we compute their probability distribution for the results presented in the previous section.

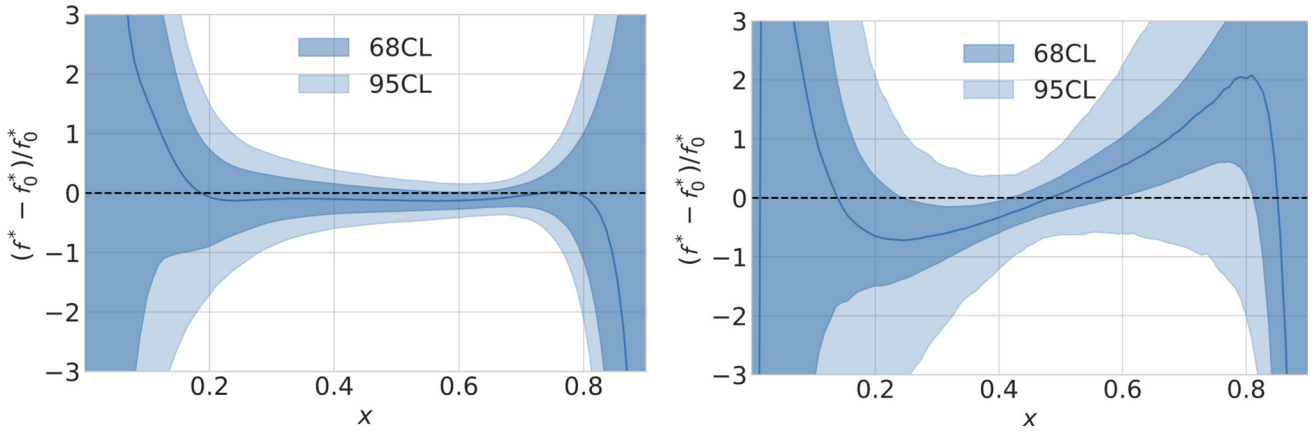

Fig. 6 Distribution of $\mathbf{f}^* - \mathbf{f}_0^* \mid (\mathrm{FK})\,\mathbf{f} + \epsilon = y$ normalized to $\mathbf{f}_0^*$

Closure tests In the context of a closure test, *i.e.* when the analysis is performed on pseudo-data built from a known model, as done in this paper, the results can be validated a posteriori, by checking how well the posterior distribution describes the underlying law. A first assessment is obtained by looking at the distribution of the stochastic variable

$$\mathbf{f}^* - \mathbf{f}_0^* \mid ((\mathrm{FK})\,\mathbf{f} + \epsilon = y)\,,$$

whose mean and covariance are given by $\tilde{\mathbf{m}}^* - \mathbf{f}_0^*$ and $\tilde{K}_{\mathbf{x}^*\mathbf{x}^*}$. In Fig. 6 we plot its distribution, normalized to $\mathbf{f}_0^*$. The left (respectively, right) panel shows the result for the case of BCDMS (respectively, lattice data): the difference is compatible with zero in the full x range, with a smaller error in the kinematic region which is more sensitive to the observed data.

In addition to the plot displayed in Fig. 6, a more quantitative measure of the inference performances can be obtained from the value of the log-loss, *i.e.* the negative logarithm of the posterior probability distribution that the stochastic variable $\mathbf{f}^*$ is equal to the underlying law $\mathbf{f}_0^*$

$$L(\mathbf{f}_0^*, \tilde{\mathbf{m}}^*, \tilde{K}_{\mathbf{x}^*\mathbf{x}^*}) = -\log\left(P\left[\mathbf{f}^* = \mathbf{f}_0^* \mid (\mathrm{FK})\,\mathbf{f} + \epsilon = y\right]\right)\,, \tag{80}$$

$$= \frac{n^*}{2}\log(2\pi) + \frac{1}{2}\log \mathrm{pdet}\,\tilde{K}_{\mathbf{x}^*\mathbf{x}^*} + \frac{1}{2}(\mathbf{f}_0^* - \tilde{\mathbf{m}}^*)^T \tilde{K}^+_{\mathbf{x}^*\mathbf{x}^*}(\mathbf{f}_0^* - \tilde{\mathbf{m}}^*). \tag{81}$$

Note that in Eq. (80), $\mathbf{f}^* \mid ((\mathrm{FK})\,\mathbf{f} + \epsilon = y)$ is a stochastic variable, whose probability density is given by the posterior Gaussian Process, while $\mathbf{f}_0^*$ is the known underlying law. For probabilities in $(0, 1)$, the log-loss is 0 when the posterior assigns probability 100% to what is actually observed, while with probability densities the scale is arbitrary. This means that in the continuous case the log-loss can be used to compare and benchmark inferences, but its absolute value does not have a definite interpretation. As one can see in Eq. (81), the log-loss penalizes not only wrong answers, *i.e.* posterior central values that differ from the underlying law, but also large errors.

Although here we wrote out the expression of the log-loss for our specific GP model (Eq. 81), the definition of log-loss is totally general within the Bayesian paradigm, so it could be used to compare any other fully Bayesian inference to our method. We do not try such a comparison in this paper.

Real data analysis When dealing with a real analysis the true underlying model is not known. Some metrics assessing a posteriori the goodness of the fit and the ability of the model to generalize to unseen data are therefore necessary to evaluate the performance of a given methodology. A possible metric is given by the quantity introduced in Eq. (24) evaluated for $\mathbf{f} = \tilde{\mathbf{m}}$, which is in some proper statistical sense the Bayesian equivalent of a more familiar frequentist χ^2:

$$\frac{S(\tilde{\mathbf{m}}; \theta, y, C_Y)}{\mathrm{dof}} = \frac{1}{N_{\mathrm{data}}}\Big((\mathbf{m} - \tilde{\mathbf{m}})^T K_{\mathbf{xx}}^{-1}(\mathbf{m} - \tilde{\mathbf{m}}) + (y - (\mathrm{FK})\tilde{\mathbf{m}})^T C_Y^{-1}(y - (\mathrm{FK})\tilde{\mathbf{m}})\Big). \tag{82}$$

The two pieces can be looked at separately to see if the fit is deviating from the prior or from the data. The usual empirical usage of expecting $S/\mathrm{dof} \approx 1$ is valid. Its distribution is plotted in Fig. 7, left panel.

When quantifying the performance of the model on a test set[5] we can compute the log-loss

$$-\log P(\mathrm{test}|\mathrm{data}).$$

[5] As customary, we denote as training data the set of data entering hyperparameter inference and gaussian conditioning, and as test data a set of out-of-sample data, *i.e.* a set of data not entering neither hyperparameter inference nor gaussian conditioning. We can think of test data as a new dataset with its own covariance C_Y^*, uncorrelated to the data used for the Bayesian inference.

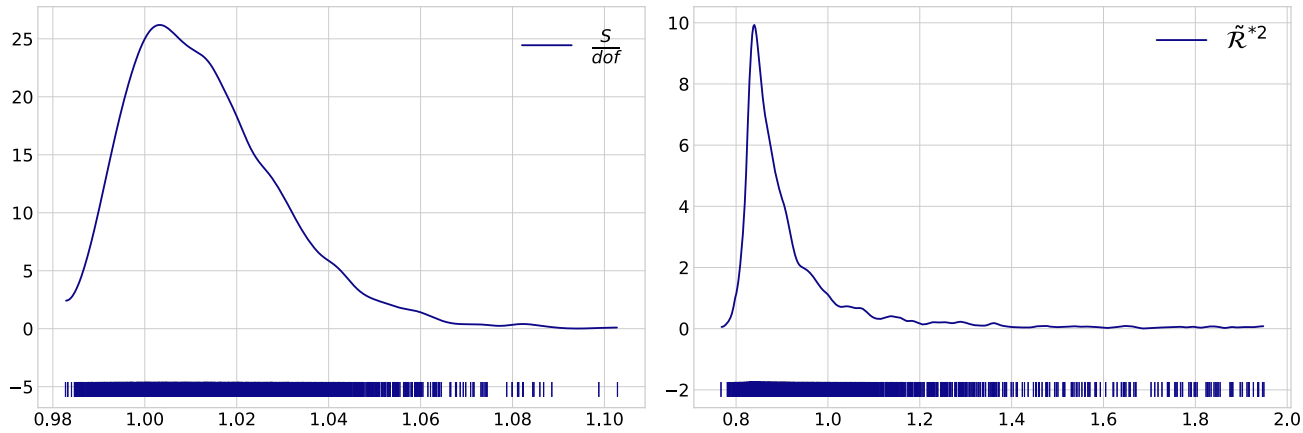

Fig. 7 S/dof (left panel) and $\tilde{\mathcal{R}}^{*2}$ (right panel). The distributions are obtained by sampling from the hyperparameter posterior $p(\theta|y, C_Y)$ and using the explicit expressions in Eqs. (82), (84)

Denoting with $*$ the quantities computed on the test points this is given by

$$L(y^*) = \frac{n^*}{2}\log(2\pi) + \frac{1}{2}\log\text{pdet}\left((\text{FK}^*)\tilde{K}_{\mathbf{x}^*\mathbf{x}^*}(\text{FK}^*)^T + C_Y^*\right) + \frac{1}{2}(y^* - (\text{FK}^*)\tilde{\mathbf{m}}^*)^T\left((\text{FK}^*)\tilde{K}_{\mathbf{x}^*\mathbf{x}^*}(\text{FK}^*)^T + C_Y^*\right)^+ \times (y^* - (\text{FK}^*)\tilde{\mathbf{m}}^*). \tag{83}$$

Excluding the first constant term, the other two terms are interpretable: the first is a log-determinant of the posterior covariance matrix, so it's a measure of the volume occupied by the distribution; in other words, it summarizes how large is the final uncertainty. The second term is the usual squared distance between prediction and data in units of the uncertainty. Having this in mind we can define the metrics

$$\tilde{\mathcal{R}}^{*2} = \frac{1}{\dim(y^*|y)}((\text{FK}^*)\tilde{\mathbf{m}} - y^*)^T \times \left((\text{FK}^*)\tilde{K}_{\mathbf{xx}}(\text{FK}^*)^T + C_Y^*\right)^+ ((\text{FK}^*)\tilde{\mathbf{m}} - y^*), \tag{84}$$

$$\tilde{\sigma}^{*2} = \exp\left(\frac{1}{\dim(y^*|y)}\log\text{pdet}\left((\text{FK}^*)\tilde{K}_{\mathbf{xx}}(\text{FK}^*)^T + C_Y^*\right)\right). \tag{85}$$

The quadratic form $\tilde{\mathcal{R}}^{*2}$ should be about 1 (to be sure that test data are described within uncertainty), while $\tilde{\sigma}$ quantifies the average posterior error in a Bayesianly justified way. By comparing the $\tilde{\sigma}$ values corresponding to different methodologies we can assess quantitatively which one is more or less conservative.

In the context of the simple example presented in this paper, we can use the BCDMS and lattice data as training and test set respectively. Using samples from $p(\theta|y, C_Y)$ and Eq. (84), we can access the full probability distribution of $\tilde{\mathcal{R}}^{*2}$, which is plotted in Fig. 7, right panel.

In this paper we do not try to compare our methodology to the standard results obtained with a non-Bayesian approach, as it would require a substantial amount of work beyond the scope of the rest of the paper. We leave it to a future analysis with a more complete DIS dataset. We just comment on how the goodness-of-fit metrics we have shown here would (or wouldn't) apply: the log-loss is defined only within a Bayesian inference, so it would not allow comparisons between our methodology and the standard ones. It would only allow comparisons, say, between different choices of kernel for the GP, or between a GP and a non-GP but still Bayesian model. On the other hand, the quantity $\mathcal{R}^{*2}$ from Eq. (84) can be generalized to a standard fit – for example, in the case of the MonteCarlo fits carried out within the NNPDF methodology, the posterior mean $\tilde{\mathbf{m}}$ and covariance matrix $\tilde{K}_{\mathbf{xx}}$ should be replaced by the mean and covariance matrix of the replicas – and would therefore allow for a quantitative comparison between our methodology and a non-Bayesian approach.

5 Summary and future work

We have described a Bayesian methodology for the solution of the inverse problem underlying the determination of PDFs. GPs are used for the modelling of the PDF prior. Known physical constraints, such as sum rules, kinematic limit and small-x power behaviour are implemented in the prior by suitable manipulation of the GP kernel. We discussed the case of observables that depend linearly on the PDF, and the analytical simplification occurring in this scenario, and we applied the methodology to two simple examples concerning the extraction of a single PDF flavor from a reduced dataset of DIS structure functions and lattice correlators. In order to validate our approach we have used pseudo-data produced from a known underlying law. We have found that, even in the pres-

ence of noisy data, the input model is reconstructed within the quoted error. We have discussed the mathematical definition of the final uncertainties given by this approach, which allows for a quantitative estimation of the different components entering the PDF error. Finally we have discussed the validation of the results by introducing a set of metrics, which allow to assess the goodness of a given methodology and compare different ones, using Bayesianly justified figures of merit.

This work is intended to be a preliminary study to explore the main features, advantages and limitations of the Bayesian approach, and it paves the way to a full PDF determination from an extended DIS dataset. This will be the object of a future separate paper, in which we aim to deliver a full DIS-only PDF set, to be compared to other available sets based on parametric regression.

In order to achieve a global PDF determination, not only based on DIS data but including also LHC hadronic observables, the general approach described here can still be applied, but no analytical simplification occur, as described in Sect. 2.5. This implies that a Monte Carlo with dimension given by the total number of parameters and hyperparameters has to be run to access the full posterior, which makes the problem computationally more expensive than the linear case, where the Monte Carlo dimension is given by the number of hyperparameters only. The general features of the Bayesian approach still hold, and the development of a framework for a global PDF determination will be the object of a further studies.

Acknowledgements TG is supported by NWO via an ENW-KLEIN-2 project. The work of LDD was supported by the ExaTEPP project EP/X01696X/1, and by the UK Science and Technology Facility Council (STFC) grant ST/P000630/1. We thank Juan Rojo and the members of the NNPDF collaboration for comments and discussions.

Data Availability Statement This manuscript has no associated data. [Authors' comment: There are no data associated with this work. No new datasets were generated or analysed for this study.]

Code Availability Statement This manuscript has associated code/software in a data repository. [Authors' comment: The code used for this study will be made available upon request.]

Funded by SCOAP[3].

Appendix A: Sum rules

Following the discussion in Sect. 2, we associate a GP to each PDF in the evolution basis. The valence distributions V_{a} obey sum rules that we want to incorporate into the prior. This can be done by associating a GP to the indefinite integral of the PDF: denoting as $\widehat{V}_{\mathsf{a}}$ the primitive of V_{a}, we associate to the former a GP having mean and kernel

$$\widehat{m}^{\mathsf{a}}(x) \quad \text{and} \quad \widehat{K}^{\mathsf{a}}(x,y)\,. \tag{86}$$

It can be shown [11] that V_{a} is then represented by a GP having as mean and kernel

$$m^{\mathsf{a}}(x) = \partial_x \widehat{m}^{\mathsf{a}}(x)\,, \quad K^{\mathsf{a}}(x,y) = \partial_x \partial_y\, \widehat{K}^{\mathsf{a}}(x,y)\,. \tag{87}$$

In formulae

$$\begin{aligned} &V_{\mathsf{a}}(x) = \widehat{V}'_{\mathsf{a}}(x)\,, \quad \widehat{V}_{\mathsf{a}} \sim \mathcal{GP}\big(0, \widehat{K}^{\mathsf{a}}(x,y)\big)\,, \\ &V_{\mathsf{a}} \sim \mathcal{GP}\big(0, K^{\mathsf{a}}(x,y)\big)\,, \quad \mathsf{a} = 1,3,8,15\,, \end{aligned} \tag{88}$$

where we used $V_1(x)$ to denote the total valence $V(x)$, the $'$ denotes the derivative with respect to x, and $K^{\mathsf{a}}(x,y)$ is given in Eq. (87). The sum rules can then be expressed by introducing additional GPs for the primitive of the PDFs, with kernels satisfying Eq. (87), and by imposing linear constraints between them. In the case of the valence sum rules we get

$$\widehat{V}(1) - \widehat{V}(0) = \widehat{V}_8(1) - \widehat{V}_8(0) = \widehat{V}_{15}(1) - \widehat{V}_{15}(0) = 3\,, \tag{89}$$

$$\widehat{V}_3(1) - \widehat{V}_3(0) = 1\,. \tag{90}$$

Similarly the momentum sum rule is written in terms of the indefinite integral of $x\Sigma$ and xg, denoted as $\widehat{x\Sigma}$ and $\widehat{xg}$, and can be imposed by introducing the GPs

$$\widehat{x\Sigma} \sim \mathcal{GP}(0, \Theta^{\Sigma}), \quad \widehat{xg} \sim \mathcal{GP}(0, \Theta^{g}), \tag{91}$$

and imposing the linear constraint

$$\widehat{x\Sigma}(1) + \widehat{xg}(1) - \widehat{x\Sigma}(0) - \widehat{xg}(0) = 1. \tag{92}$$

The power behaviour for $x \to 0^+$ of a given PDF can be enforced by rescaling the corresponding kernel function according to Eq. (53). In the case in which sum rules and small-x power behaviour need to be imposed at the same time, the rescaling function of Eq. (54) should be applied at the kernel representing the primitive. For example, if we want te valence distribution V to scale as x^{α_V}, then the kernel for $\widehat{V}$ should be rescaled using

$$\phi(x) = x^{\alpha_V + 1}.$$

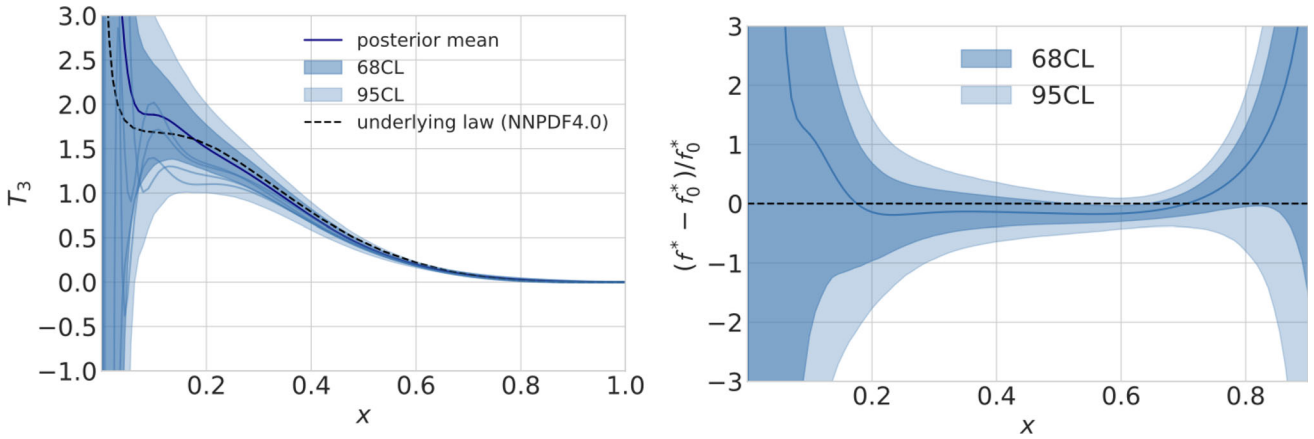

Fig. 8 Samples from $p(\mathbf{f}^*, \alpha, l_0, \sigma | y, C_Y)$ plotted in linear scale (left panel) and distribution of $\mathbf{f}^* - \mathbf{f}_0^* \mid (\mathrm{FK})\,\mathbf{f} + \epsilon = y$ normalized to $\mathbf{f}_0^*$ (right panel). Both plots refer to the case of the analysis on the BCDMS data accounting for the kinetic constrain $T_3(1) = 0$

Since for the primitives we use a Gibbs kernel (Eq. (50)) with variable length scale, the length scale affects the amplitude of the derivatives. The length scale is proportional to x, so the derived process has standard deviation proportional to $1/x$. In our specific case, it turns out that the correction does not alter the intended variance of the derived process: $\partial(\phi(x)\, f(x)) \sim x^{\alpha_V} \cdot 1 + x^{\alpha_V+1} \cdot x^{-1} \sim x^{\alpha_V}$.

Appendix B: Change of the prior in the extrapolation region: kinetic limit

As pointed out in Sect. 3.4, the behavior of the posterior in the extrapolation regions strongly depends on the specific choice we make for the prior: in the absence of any experimental information the posterior reduces to the prior. The kinematic constraint according to which all flavors vanish at $x = 1$, known as kinetic limit, is an example of a property that, when implemented, directly modifies the prior in the large-x extrapolation region. Given that the conditions

$$\Sigma(1) = g(1) = T_{\mathsf{a}}(1) = V(1) = V_{\mathsf{a}}(1) = 0, \quad \mathsf{a} = 3, 8, 15, \tag{93}$$

are simple linear constraints involving each individual flavor, they can be implemented in the prior by treating them as additional datapoints, extending the FK table. In the left panel Fig. 8 we show the results we got for the posterior when performing the analysis on the BCDMS data accounting for the kinetic limit: unlike the analogous plot in Fig. 2 – where at large x the error increases reflecting the lack of experimental data – the error in $x = 1$ now shrinks to 0, according to the new information we implanted in the prior.

In the right panel of Fig. 8 we plot the distribution of

$$\mathbf{f}^* - \mathbf{f}_0^* \mid ((\mathrm{FK})\,\mathbf{f} + \epsilon = y)\,,$$

which allows to check that the additional constraint is not introducing a bias. Comparing this result with the analogous plot in Fig. 6, it is clear how by imposing the kinetic limit we obtain a better description of the underlying law at large-x.

References

1. S. Bailey, T. Cridge, L.A. Harland-Lang, A.D. Martin, R.S. Thorne, Parton distributions from LHC, HERA, Tevatron and fixed target data: MSHT20 PDFs. Eur. Phys. J. C **81**(4), 341 (2021). arXiv:2012.04684
2. T.-J. Hou et al., New CTEQ global analysis of quantum chromodynamics with high-precision data from the LHC. Phys. Rev. D **103**(1), 014013 (2021). arXiv:1912.10053
3. **NNPDF** Collaboration, R.D. Ball et al., The path to proton structure at 1% accuracy. Eur. Phys. J. C **82**(5), 428 (2022). arXiv:2109.02653
4. **H1, ZEUS** Collaboration, H. Abramowicz, et al., Combination of measurements of inclusive deep inelastic $e^{\pm} p$ scattering cross sections and QCD analysis of HERA data. Eur. Phys. J. C **75**(12), 580 (2015). arXiv:1506.06042
5. **ATLAS** Collaboration, G. Aad et al., A precise determination of the strong-coupling constant from the recoil of Z bosons with the ATLAS experiment at $\sqrt{s} = 8$ TeV, arXiv:2309.12986
6. **ATLAS** Collaboration, Improved W boson Mass Measurement using 7 TeV Proton-Proton Collisions with the ATLAS Detector, tech. rep., CERN, Geneva (2023). All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2023-004
7. **PDF4LHC Working Group** Collaboration, R. D. Ball et al., The PDF4LHC21 combination of global PDF fits for the LHC Run III. J. Phys. G **49**(8), 080501(2022). arXiv:2203.05506
8. R.D. Ball, , V. Bertone, S. Carrazza, C.S. Deans, L.D. Debbio, S. Forte, A. Guffanti, N.P. Hartland, J.I. Latorre, J. Rojo, M. Ubiali, Parton distributions for the LHC run II. J. High Energy Phys. **2015** (2015)
9. L. Del Debbio, T. Giani, M. Wilson, Bayesian approach to inverse problems: an application to NNPDF closure testing. Eur. Phys. J. C **82**(4), 330 (2022). arXiv:2111.05787
10. A. Candido, L. Del Debbio, T. Giani, G. Petrillo, Inverse problems in PDF determinations. PoS **LATTICE2022**, 098 (2023). arXiv:2302.14731

11. C.E. Rasmussen, C.K.I. Williams, *Gaussian Processes for Machine Learning, Adaptive Computation and Machine Learning* (MIT Press, Cambridge, 2006)
12. A.P. Valentine, M. Sambridge, Gaussian process models-I. A framework for probabilistic continuous inverse theory. Geophys. J. Int. **220** 1632–1647 (2019). https://academic.oup.com/gji/article-pdf/220/3/1632/31578341/ggz520.pdf
13. M. Hansen, A. Lupo, N. Tantalo, Extraction of spectral densities from lattice correlators. Phys. Rev. D **99**(9), 094508 (2019). arXiv:1903.06476
14. J. Horak, J.M. Pawlowski, J. Rodríguez-Quintero, J. Turnwald, J.M. Urban, N. Wink, S. Zafeiropoulos, Reconstructing QCD spectral functions with Gaussian processes. Phys. Rev. D **105**(3), 036014 (2022). arXiv:2107.13464
15. J. Karpie, K. Orginos, A. Rothkopf, S. Zafeiropoulos, Reconstructing parton distribution functions from Ioffe time data: from Bayesian methods to neural networks. JHEP **04**, 057 (2019). arXiv:1901.05408
16. J. Horak, J.M. Pawlowski, J. Turnwald, J.M. Urban, N. Wink, S. Zafeiropoulos, Nonperturbative strong coupling at timelike momenta. Phys. Rev. D **107**(7), 076019 (2023). arXiv:2301.07785
17. J.R. Schott, *Matrix Analysis for Statistics*, 3rd edn. (Wiley, New York, 2017)
18. K.P. Murphy, *Probabilistic Machine Learning: Advanced Topics* (MIT Press, Cambridge, 2023)
19. B.C.D.M.S. Collaboration, A.C. Benvenuti et al., A high statistics measurement of the proton structure functions $F_2(x, Q^2)$ and R from deep inelastic muon scattering at high Q^2. Phys. Lett. B **223**, 485 (1989)
20. A.-P. Oriol, A. Virgile, C. Colin, D. Larry, F.C. J., K. Maxim, K. Ravin, L. Jupeng, L. C. C., M. O. A., O. Michael, V. Ricardo, W. Thomas, Z. Robert, Pymc: a modern and comprehensive probabilistic programming framework in python. PeerJ Comput. Sci. **9**, e1516 (2023)
21. K. Cichy, L. Del Debbio, T. Giani, Parton distributions from lattice data: the nonsinglet case. JHEP **10**, 137 (2019). arXiv:1907.06037
22. L. Del Debbio, T. Giani, J. Karpie, K. Orginos, A. Radyushkin, S. Zafeiropoulos, Neural-network analysis of parton distribution functions from Ioffe-time pseudo distributions. JHEP **02**, 138 (2021). arXiv:2010.03996
23. A.V. Radyushkin, Quasi-parton distribution functions, momentum distributions, and pseudo-parton distribution functions. Phys. Rev. D **96**(3), 034025 (2017). arXiv:1705.01488