PROCEEDINGS OF SCIENCE

Markov Chain Monte Carlo techniques applied to Parton Distribution Functions determination: proof of concept

PoS(DIS2017)213

Mariane Mangin-Brinet*, **Yémalin Gabin Gbedo**
Laboratoire de Physique Subatomique et de Cosmologie - Université Grenoble-Alpes,
CNRS/IN2P3, 53, avenue des Martyrs, 38026 Grenoble, France
E-mail: mariane@lpsc.in2p3.fr, gbedo@lpsc.in2p3.fr

We propose a Bayesian parameter inference approach to determine Parton Distribution Functions (PDFs) and we show that we can replace the standard χ^2 minimisation used in most existing PDF global analysis procedures, by Markov chain Monte Carlo (MCMC) techniques. These methods, widely used in statistics, lead to reliable estimates of uncertainties in terms of confidence limit intervals of probability distributions, and offer additional insight into the rich field of PDFs. The formulation of PDF determination in terms of Bayesian inference, the Monte Carlo algorithm we have implemented in the xFitter code and a selection of first results we have obtained are presented in this contribution.

XXV International Workshop on Deep-Inelastic Scattering and Related Subjects
3-7 April 2017
University of Birmingham, UK

*Speaker.

1. Introduction and motivation

Parton Distribution Functions (PDFs) are fundamental ingredients of QCD, and with the advent of the Large Hadron Collider – powerful QCD factory operating in an unexplored energy range, the need to determine PDFs and to assess their associated uncertainties as precisely as possible has become crucial. Except neural network techniques, most existing procedures to assess PDF uncertainties rely on a minimisation procedure and on the choice of a tolerance criteria to define the permissible range of "acceptable" $\Delta\chi^2$ one has to explore around the minimal χ^2 [1]. Uncertainties estimated this way thus lose their statistical meaning. To define the uncertainties in a way based as much as possible on robust statistical methods, we propose to use Bayesian parameter inference and Markov chain Monte Carlo (MCMC) techniques, which have been an extremely popular tool in statistics. These methods allow to estimate *a posteriori* probability densities for multi-dimensional models and provides reliable estimates of errors.

In the next section of this paper, we show how to formulate the PDF determination problem in terms of Bayesian inference. We then recall in the third section basic principles of Markov chain Monte Carlo methods, explain which algorithm we have chosen to implement and why, and briefly sketches the procedure to extract relevant informations from Markov chains. Section 4 present some preliminary results.

2. Formulation of the PDF determination in terms of Bayesian inference

Parton Distribution Function determination in the context of global analysis consists in extensively exploitating of datasets collected at colliders to constrain the parameters of the PDF functional forms given at a fixed scale in energy. For compactness, let us note $\hat{q}$ the vector of PDF parameters to be determined: $\hat{q} = (q^{(1)}, q^{(2)}, \ldots, q^{(m)})^T$ where m is typically, in the case of a full analysis, of the order of 25-30, and D the data. From a Bayesian perspective, both model parameters $\hat{q}$ and observables are considered random quantities, and Bayesian inference aims at the determination of the distribution of the parameters $\hat{q}$ conditional on the data D : $P(\hat{q}|D)$. This so-called *posterior* probability density, which quantifies the probability to have the model parameters $\hat{q}$ given the observed data D, is expressed by Bayes theorem in terms of the likelihood $P(D|\hat{q}) \overset{\text{def}}{=} \mathscr{L}(D,\hat{q})$ by :

$$P(\hat{q}|D) = \frac{\mathscr{L}(D,\hat{q})P(\hat{q})}{\int d\hat{q}\mathscr{L}(D,\hat{q})P(\hat{q})} \tag{2.1}$$

where $P(\hat{q})$ is a *prior* distribution, quantifying the degree of belief one has *a priori* before observing the data and the denominator can be considered only as a normalization.

To determine this conditional probability, we thus need to set a prior distribution for the parameters, and to compute the likelihood of the data. The probability density $P(\hat{q}|D)$ is then sampled using a Monte Carlo algorithm.

Using the fact that the least square method and the maximum likelihood should be equivalent in the case of normally distributed data, we construct the likelihood of the data in the same way the χ^2 is defined, and we identify $\log \mathscr{L}(D,\hat{q}) = -\frac{1}{2}\chi^2$. In the feasibility study we present, the χ^2 does not include any correlation, but more generally, correlated experimental uncertainties can be taken into account by introducing for instance a covariant matrix and properly modifying the χ^2.

3. Markov chain Monte Carlo in a nutshell

MCMC algorithms enable us to draw samples from a probability distribution known up to a multiplicative constant, and consist in sequentially simulating a single Markov chain whose limiting distribution is the chosen one (in our case, the maximum likelihood times a prior density). Basic ingredients of MCMC are illustrated in the following section using the Metropolis algorithm.

3.1 Basic principles of Markov chains and Metropolis algorithm

Two ingredients are necessary to define a Markov chain: (i) the initial values (that is the marginal distribution) of parameters and (ii) the transition kernel between two sets of parameters: $T(\hat{q} \longrightarrow \hat{q}')$, for going from a set $\hat{q}$ to another set $\hat{q}'$. The standard computational workhorse of MCMC methods is the so called "Metropolis-Hastings algorithm", proposed in 1953 by Metropolis et al. and generalized by Hastings in 1970 [2]. It can be applied in principle to any system and is extremely straightforward to implement. It proceeds as follows: at each Monte Carlo time $t-1$, the next state $\hat{q}_t$ is chosen by sampling a candidate point $\hat{q}'$ from a proposal distribution $\pi(.|\hat{q}_{t-1})$. The candidate point is then accepted with the probability

$$\alpha(\hat{q}_{t-1}, \hat{q}') = \min\left(1, \frac{P(\hat{q}'|D)\pi(\hat{q}_{t-1}|\hat{q}')}{P(\hat{q}_{t-1}|D)\pi(\hat{q}'|\hat{q}_{t-1})}\right)$$

and the Metropolis-Hastings transition kernel is thus

$$T(\hat{q}_{t-1} \longrightarrow \hat{q}') = \pi(\hat{q}'|\hat{q}_{t-1})\alpha(\hat{q}_{t-1}, \hat{q}').$$

If the new set of parameters $\hat{q}'$ is accepted, the next state of the chain becomes $\hat{q}_t = \hat{q}'$. If it is rejected, the chain does not move and the point at t is identical to the point at $t-1$: $\hat{q}_t = \hat{q}_{t-1}$. The main drawback of this algorithm is the fact that the autocorrelations become large and the acceptance very tiny as the dimension of the parameter space to explore increases. For realistic PDF determination, where the number of parameters can be of the order of a several dozens, Metropolis algorithm – even improved by techniques like multivariate Gaussian distributions, binary space partitioning ...– is inefficient. This is the reason why we have implemented a much more elegant algorithm, based on Molecular Dynamics, which has initially been developed for Lattice QCD and is widely used in this field.

3.2 Hybrid Monte Carlo algorithm

Hamiltonian (or "hybrid") dynamics [3], developed originally for lattice field theory, is used to produce candidate proposals for Metropolis algorithm, in a very elegant and efficient way. It is an exact algorithm which combines molecular dynamics evolution with a Metropolis accept/reject step, which is used to correct for discretization errors in the numerical integration of the corresponding equations of motion.

Hybrid Monte Carlo consists in associating to each set of parameters $\hat{q}$ (see previous section) a set of conjugate momenta $\hat{p}$ and to replace the *a posteriori* probability density (2.1) we want to sample by the joint distribution defined as

$$P(\hat{q}, \hat{p}) = \frac{1}{Z}e^{-H(\hat{q},\hat{p})} = \frac{1}{Z}e^{-\mathscr{K}(\hat{p})}e^{-\mathscr{U}(\hat{q})}$$

where Z is a normalizing constant and $H(\hat{q},\hat{p})$ is an hamiltonian written as $H(\hat{q},\hat{p}) = \mathscr{K}(\hat{p}) + \mathscr{U}(\hat{q})$. The first term has the form of a kinetic energy $\mathscr{K}(\hat{p}) = \hat{p}^T M^{-1}\hat{p}/2$, where M is a mass matrix (generally taken to be diagonal) and $\mathscr{U}(\hat{q})$ is an arbitrary potential energy, that we define as $\mathscr{U}(\hat{q}) = -\log[\mathscr{L}(D,\hat{q})P(\hat{q})]$.

Starting from a point $\hat{q}_0$ of the chain, the HMC procedure consists in selecting some initial momenta $\hat{p}_0$ normally distributed around zero and let the system evolve deterministically for some time according to Hamilton's equations of motion for $H(\hat{q},\hat{p})$. It reaches a candidate point $(\hat{q}_1,\hat{p}_1)$ which, according to Metropolis procedure described above, is accepted with probability $min(1,e^{-\Delta H})$. Since the dynamics conserves energy, i.e. $\Delta H = 0$ along a trajectory, the acceptance rate is 100%, independently of the dimension of the vector $\hat{q}$. Even though in practice, the acceptance is degraded because of the numerical resolution of Hamilton equations, it can be kept very high (typically of the order of 70-90%), independently of the dimension of the chain, that is of the number of parameters to determine. More details about this algorithm can be found in many very good reviews and papers and we refer the reader for instance to [4] and references therein.

3.3 Markov chain analysis

Extracting observables and assessing their statistical errors in Monte Carlo simulations is far from begin a trivial task and it requires a careful treatment of the Markov chain. We will only briefly sketch in what follows, the procedure we have used; more details and results can be found in [5].

Among all the subtleties of Markov chain analysis, two points in particular deserve special care: the determination of the thermalization region and the estimation of the autocorrelation time.

The thermalization time of a Markov chain corresponds to a number of states to be discarded from the beginning so that the chain forgets its starting point; we have estimated it on a criteria based on the median value of the target distribution $P(\hat{q}|D)$.

To estimate the autocorrelation time, we have used either the Γ-method [6], which consists in explicitly determining autocorrelation functions and the autocorrelation time τ_{int}, or a subsampling consisting in rejecting all states which are closer than $2\tau_{int}$ to each other, in order to get independent states. We estimated errors on the uncorrelated measurements obtained with this latter method, by the Jackknife binning procedure [7]. We have checked that both methods give coherent results.

We present in the next section a selection of preliminary results obtained after skipping thermalization and properly taking into account the autocorrelation. More results can be found in [5].

4. Preliminary results

We have implemented the Hybrid Monte Carlo algorithm in the open-source package HeraFitter and its successor xFitter [8]. We use for the PDF parametrization, the HERAPDF functional form, that we just recall here for the sake of clarity: the parametrized HERA PDFs are the valence distribution xu_{val} and xd_{val}, the gluon distribution xg, and the $\overline{U}$ and $\overline{D}$ distribution defined as $x\overline{U} = x\overline{u}$, $x\overline{D} = x\overline{d} + x\overline{s}$. Their functional form at the initial scale $Q_0 = 1.9$ GeV^2 reads $xf_a(x) = A_a x^{B_a}(1-x)^{C_a}(1 + D_a x + E_a x^2)$ where a labels a parton (g, u_{val}, d_{val}, ... See [8] for more details).

The results shown in this section are obtained from 36 Monte Carlo chains – each starting from a different random point – using the HERAPDF functional forms at a scale $Q_0 = 1.9$ GeV^2 with 10 free parameters: B_g, C_g, $B_{u_{val}}$, $C_{u_{val}}$, $E_{u_{val}}$, $C_{d_{val}}$, $C_{\overline{U}}$, $A_{\overline{D}}$, $B_{\overline{D}}$ and $C_{\overline{D}}$ (see [8] for more details). We have used uniform priors for the parameters, and we consider the same data ensembles than the ones used to produce HERAPDF1.0 distributions, with the exception of the heavy flavor scheme, which is in our case the ZMVFN scheme. All chains are thermalized and decorrelated according to the procedure mentioned in the previous section and detailed in [5].

To extract parton distribution functions from the Markov chain, we compute, from the set of 10 parameters obtained at each Monte Carlo iteration, the corresponding PDFs for a range of x and Q^2 values. This provides the marginal probability density functions of PDFs at fixed (x, Q^2), as illustrated on Figure 1 for the gluon, for two different x values.

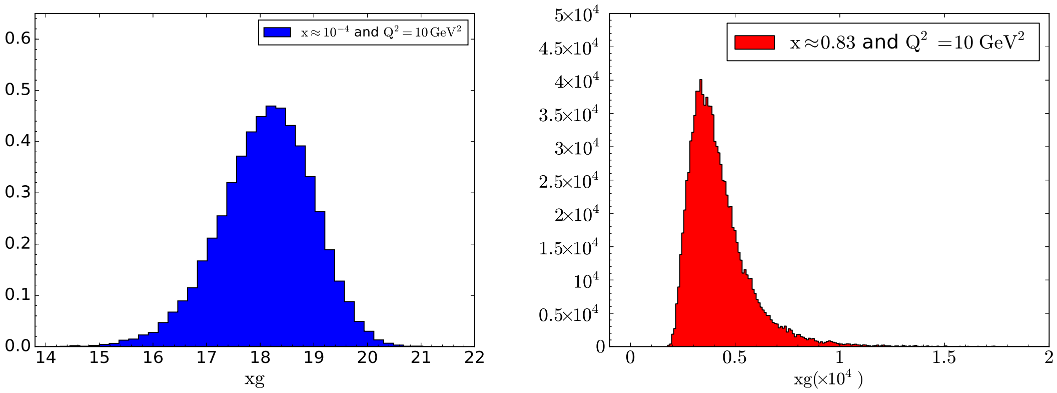

Figure 1: *Gluon PDF probability distribution function for $x \approx 10^{-4}$ (l.h.s.) and $x \approx 0.83$ (r.h.s.) at fixed $Q^2 = 10$ GeV^2. The 68% confidence interval is obtained from this distribution, considering the region of the distribution containing 68% of the data remaining on each side of the best fit value.*

For each (x, Q^2), we determine the α%-confidence interval around the best fit value of the PDF (with typically $\alpha = 68$ or $\alpha = 95$) by considering the region of the distribution on each side of the best fit, and taking α% of the data on each of these regions.

MCMC PDFs are found to be, as expected, very close to the HERAPDF1.0 PDFs (in ZMVFN scheme), both in central value and in confidence interval. Maximum likelihood estimator and least square method are indeed equivalent under Gaussian assumption, which in the case of HERAPDF1.0 settings, can be reasonably applied. Experimental uncertainties – normalized by the best fit value – obtained for HERAPDF1.0 and MCMC respectively by the Hessian and MCMC methods are compared in Fig. 2 for the u_{val} and the gluon distributions. They are consistent within the kinematic range of HERA, even if MCMC uncertainties tend to be slightly larger than the standard deviations obtained in the Hessian approach. These results validate our implementation of MCMC in xFitter and pave the way for a more complete PDF determination by MCMC techniques.

5. Conclusion and outlook

We have shown that Bayesian parameter inference approach applied to global PDFs analysis can lead to a deeper insight into PDFs uncertainties. The innovative procedure we have implemented, which combines Monte Carlo techniques, lattice-developed algorithms and global PDFs

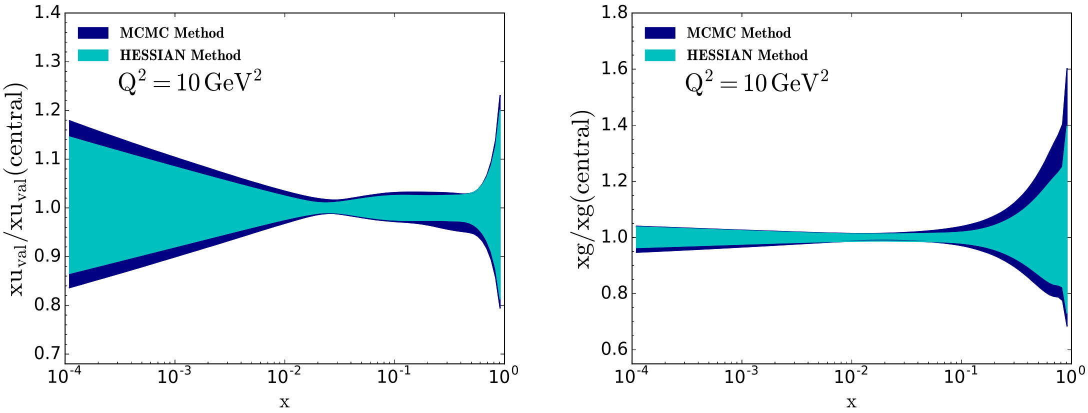

Figure 2: *Comparison of the PDF uncertainties, normalized by the best fit value, as determined by the Hessian and MCMC methods at NLO for the valence distribution xu_{val} and the gluon distribution xg, at a scale* $\mathrm{Q}^2 = 10\ \mathrm{GeV}^2$.

analysis is complementary to the existing methods. Our goal is to extend the present work to the full ensemble of PDF free parameters, including also as parameters, the strong coupling constant and c and b quark masses. We will consider more complex χ^2 functions including correlation and complete our analysis on a fully realistic case, studying in particular the impact of priors. No doubt that Markov Chain Monte Carlo methods will give interesting and valuable informations on PDFs and will contributed to our deeper understanding of these key elements of QCD.

6. Acknowledgments

We are grateful to the CFM Foundation for Research for its financial support through the Jean-Pierre Aguilar 2014 grant. Most of the computation has been done in Lyon-CCIN2P3, thanks to the CPUs resources provided. We thank the staff members of this computing center for their support.

References

[1] A. D. Martin, R. G. Roberts, W. J. Stirling and R. S. Thorne, Eur. Phys. J. **C 28**, 455 (2003); **C35**, 325 (2004); W. T. Giele and S. Keller, Phys. Rev. **D58**, 094023 (1998); J. Pumplin, D. Stump, R. Brock, D. Casey, J. Huston, J. Kalk, H.L. Lai, W.K. Tung, Phys. Rev. **D65**, 014013 (2001); D. Stump, J. Pumplin, R. Brock, D. Casey, J. Huston, J. Kalk, H.L. Lai, W. K. Tung, Phys. Rev. **D65**, 014012 (2001).

[2] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem Phys. **21**, 1087 (1953); W. K. Hastings, Biometrika, **57**, No. 1 (Apr, 1970), 97-109.

[3] S. Duane, A. D. Kennedy, B. J. Pendleton and D. Roweth, Phys. Lett. **B195**, 216 (1987).

[4] R. M. Neal, in *Handbook of Markov Chain Monte Carlo*, edited by S. Brooks, A. Gelman, G. L. Jones and X. Meng (Chapman and Hall/CRC, London, 2011).

[5] Y. G. Gbedo and M. Mangin-Brinet, arXiv:1701.07678 , to be published in Phys. Rev. D (2017).

[6] U. Wolff, Comput. Phys. Commun. **156**, 143 (2004).

[7] M. H. Quenouille, Biometrika, **Vol. 43**, No 3/4, 353 (1956).

[8] S. Alekhin et al., Eur. Phys. J. **C75** No 7, 304 (2015), http://www.xfitter.org/.