Bayesian perspective on global analysis of PDFs and FFs

Nobuo Sato
University of Connecticut/JLab
Seminar at MSU
MSU, 2017

Jefferson Lab Angular Momentum Collaboration

Jefferson Lab
EXPLORING THE NATURE OF MATTER

Outline

- Statistics and fitting methodology
- Applications

The parent distribution

"If we could make an infinite number of measurements, then we could describe exactly the distribution of the data points. This is not possible in practice, but we can hypothesize the existence of such a distribution that determines the probability of getting any particular observation in a single measurement. This distribution is called **parent distribution**. Similarly we can hypothesize that the measurements we have make are samples from the parent distribution and they form the sample distribution. In the limit of an infinite number of measurements, the sample distribution becomes the parent distribution"

Data reduction and error analysis for the physical sciences
Bevington and Robison

Bayesian perspective for global fits

- Consider a quantity f for which we want to determine its parent distribution

$$\mathcal{P}(f)$$

- We are interested in the case where f cannot be measured directly, but instead it is inferred from experimental data. In this case the parent distribution is conditioned to the evidence, and mathematically this is written as

$$\mathcal{P}(f|data)$$

- How do we compute $\mathcal{P}(f|data)$?
  $\rightarrow$ Bayes theorem:

$$\mathcal{P}(f|data) = \frac{1}{Z}\mathcal{L}(data|f)\pi(f)$$

$\mathcal{L}(data|f)$: Likelihood
$\pi(f)$: prior
Z: evidence

Bayesian perspective for global fits

- The likelihood function is chosen to describe the probability of the data to be drawn from a model with a given f. e.g *Gaussian likelihood*

$$\mathcal{L}(data|f) = \exp\left[-\frac{1}{2}\sum_i \left(\frac{d_i - \mathrm{model}_i(f)}{\delta d_i}\right)^2\right]$$

T. Bayes.

- The prior function allows us to restrict unphysical regions of f. We make the priors to be as flat as possible to avoid biases (uninformative priors) i.e.

$$\pi(f) = \begin{cases} 1 & \mathrm{condition}(f) == \mathrm{True} \\ 0 & \mathrm{condition}(f) == \mathrm{False} \end{cases}$$

$$\boxed{\mathcal{P}(f|d) = \frac{1}{Z}\mathcal{L}(d|f)\pi(f)}$$

Bayesian perspective for global fits

- In practice f needs to be represented mathematically e.g

$$f(x) = Nx^a(1-x)^b(1+c\sqrt{x}+dx+...)$$
$$f(x) = Nx^a(1-x)^b\mathrm{NN}(x;\{w_i\})$$
$$f(x) = \mathrm{NN}(x;\{w_i\}) - \mathrm{NN}(1;\{w_i\})$$

- The parent distribution for f becomes

$$\boldsymbol{a} = (N, a, b, c, d, ...)$$
$$\mathcal{P}(\boldsymbol{a}|d) = \frac{1}{Z}\mathcal{L}(d|\boldsymbol{a})\pi(\boldsymbol{a})$$
$$\mathcal{L}(d|\boldsymbol{a}) = \exp\left[-\frac{1}{2}\sum_i\left(\frac{d_i - \mathrm{model}_i(\boldsymbol{a})}{\delta d_i}\right)^2\right]$$
$$\pi(\boldsymbol{a}) = \prod_i \theta(a_i - a_i^{min})\theta(a_i^{max} - a_i)$$

$$\boxed{\begin{array}{c}\mathcal{P}(f|d) = \dfrac{1}{Z}\mathcal{L}(d|f)\pi(f) \\ \downarrow \\ \mathcal{P}(\boldsymbol{a}|d) = \dfrac{1}{Z}\mathcal{L}(d|\boldsymbol{a})\pi(\boldsymbol{a})\end{array}}$$

Bayesian perspective for global fits

- Having the parent distribution we can compute

$$\mathrm{E}[\mathcal{O}] = \int d^n a \;\; \mathcal{P}(\boldsymbol{a}|data) \;\; \mathcal{O}(\boldsymbol{a})$$

$$\mathrm{V}[\mathcal{O}] = \int d^n a \;\; \mathcal{P}(\boldsymbol{a}|data) \;\; (\mathcal{O}(\boldsymbol{a}) - \mathrm{E}[\mathcal{O}])^2$$

- $\mathcal{O}$ is any function of $\boldsymbol{a}$. e.g

$$\mathcal{O}(\boldsymbol{a}) = f(x; \boldsymbol{a})$$

$$\mathcal{O}(\boldsymbol{a}) = \int_x^1 \frac{d\xi}{\xi} C(\xi) f\left(\frac{x}{\xi}; \boldsymbol{a}\right)$$

- How do we compute $\mathrm{E}[\mathcal{O}], \mathrm{V}[\mathcal{O}]$?
  - **Maximum likelihood**
  - **Monte Carlo approach**

Attention:

- typically $n \gg 1$
- $\mathcal{P}(\boldsymbol{a}|data)$ is computationally expensive
- for $\mathcal{O} == f(x)$, an n–dim integration is needed for each x. Not practical!

Maximum Likelihood

- Estimation of expectation value

$$\mathrm{E}[\mathcal{O}] = \int d^n a \;\; \mathcal{P}(\boldsymbol{a}|data) \;\; \mathcal{O}(\boldsymbol{a}) \simeq \mathcal{O}(\boldsymbol{a}_0)$$

- $\boldsymbol{a}_0$ is estimated from optimization algorithm

$$\max\left[\mathcal{P}(\boldsymbol{a}|data)\right] = \mathcal{P}(\boldsymbol{a}_0|data)$$
$$\max\left[\mathcal{L}(data|\boldsymbol{a})\pi(\boldsymbol{a})\right] = \mathcal{L}(data|\boldsymbol{a}_0)\pi(\boldsymbol{a}_0)$$

- equivalently

$$\min\left[-2\log\left(\mathcal{L}(data|\boldsymbol{a})\pi(\boldsymbol{a})\right)\right] = -2\log\left(\mathcal{L}(data|\boldsymbol{a}_0)\pi(\boldsymbol{a}_0)\right)$$
$$= \sum_i \left(\frac{d_i - \mathrm{model}_i(\boldsymbol{a}_0)}{\delta d_i}\right)^2 - 2\log\left(\pi(\boldsymbol{a}_0)\right)$$
$$= \chi^2(\boldsymbol{a}_0) - 2\log\left(\pi(\boldsymbol{a}_0)\right)$$

this is Chi-squared minimization

Maximum Likelihood + Hessian method

- Estimation of variance

$$\mathrm{V}[\mathcal{O}] = \int d^n a \ \ \mathcal{P}(\boldsymbol{a}|data) \ \ (\mathcal{O}(\boldsymbol{a}) - \mathrm{E}[\mathcal{O}])^2$$

- Eigen direction decomposition of $\mathcal{P}(\boldsymbol{a}|data)$

$$\begin{aligned}
\mathcal{P}(\boldsymbol{a}|data) &\propto \exp\left(-\frac{1}{2}\chi^2(\boldsymbol{a})\right) \propto \exp\left(-\frac{1}{2}\chi^2(\boldsymbol{a}_0) - \frac{1}{2}\Delta\chi^2(\boldsymbol{a})\right) \\
&\propto \exp\left(-\frac{1}{2}\Delta\chi^2(\boldsymbol{a})\right) \\
&\propto \exp\left(-\frac{1}{2}\Delta\boldsymbol{a}^T\, H\, \Delta\boldsymbol{a}\right) + O(\Delta a^3) \\
&\propto \exp\left(-\frac{1}{2}\sum_k \left(t_k \frac{\hat{\boldsymbol{e}}_k^T}{\sqrt{w_k}}\right) H \sum_l \left(t_l \frac{\hat{\boldsymbol{e}}_l}{\sqrt{w_l}}\right)\right) + O(\Delta a^3) \\
&\propto \exp\left(-\frac{1}{2}\sum_k t_k^2\right) + O(\Delta a^3) \\
&\propto \prod_k \exp\left(-\frac{1}{2}t_k^2\right) + O(\Delta a^3)
\end{aligned}$$

$$\boxed{H\hat{\boldsymbol{e}}_k = w_k \hat{\boldsymbol{e}}_k}$$

The probability distribution “factorizes” along each eigen direction

Maximum Likelihood + Hessian method

- Estimation of variance

$$\mathrm{V}[\mathcal{O}] = \int d^n a \;\; \mathcal{P}(\boldsymbol{a}|data) \;\; (\mathcal{O}(\boldsymbol{a}) - \mathrm{E}[\mathcal{O}])^2$$

- Linear approximation of $\mathcal{O}(\boldsymbol{a})$

$$[\mathcal{O}(\boldsymbol{a}) - \mathrm{E}[\mathcal{O}]]^2 = \left[\sum_i \frac{\partial \mathcal{O}}{\partial a_i}(a_i - a_0) + O(a^2)\right]^2 = \left[\sum_k \frac{\partial \mathcal{O}}{\partial t_k} t_k\right]^2 + O(a^3)$$

- Combining with factorized $\mathcal{P}(\boldsymbol{a}|data)$ we get

master formulas

$$\mathrm{V}[\mathcal{O}] \simeq \prod_k \int dt_k \frac{e^{-\frac{1}{2}t_k^2}}{\sqrt{2\pi}} \sum_{lm} \frac{\partial \mathcal{O}}{\partial t_l} \frac{\partial \mathcal{O}}{\partial t_m} t_l t_m$$

$$= \sum_k \left(\frac{\partial \mathcal{O}}{\partial t_k}\right)^2 \simeq \sum_k \left[\frac{\mathcal{O}(t_k = 1) - \mathcal{O}(t_k = -1)}{2}\right]^2$$

Maximum Likelihood + Hessian method

- **pros**

$\rightarrow$ Very practical. Most the PDF groups use this method

$\rightarrow$ It is computationally inexpensive

$\rightarrow$ f and its eigen directions can be precalculated/tabulated

- **cons**

$\rightarrow$ Assumes local gaussian approximation of the likelihood

$\rightarrow$ Assumes linear approximation of the observables $\mathcal{O}$ around $\boldsymbol{a}_0$

$\rightarrow$ The assumptions are strictly valid for linear models.

$\rightarrow$ Computation of the hessian matrix is numerically unstable if flat directions are present

- **examples**

$\rightarrow$ if $f(x) = a + bx + cx^2$ then $\mathrm{E}[f(x)] = \mathrm{E}[a] + \mathrm{E}[b]x + \mathrm{E}[c]x^2$

$\rightarrow$ but $f(x) = Nx^a(1-x)^b$ then $\mathrm{E}[f(x)] \neq \mathrm{E}[N]x^{\mathrm{E}[a]}(1-x)^{\mathrm{E}[b]}$

Monte Carlo Methods

- Recall that we are interested in computing

$$\mathrm{E}[\mathcal{O}] = \int d^n a \;\; \mathcal{P}(\boldsymbol{a}|data) \;\; \mathcal{O}(\boldsymbol{a})$$

$$\mathrm{V}[\mathcal{O}] = \int d^n a \;\; \mathcal{P}(\boldsymbol{a}|data) \;\; (\mathcal{O}(\boldsymbol{a}) - \mathrm{E}[\mathcal{O}])^2$$

- Any MC method attempts to do this using MC sampling

$$\mathrm{E}[\mathcal{O}] \simeq \sum_k w_k \mathcal{O}(\boldsymbol{a}_k)$$

$$\mathrm{V}[\mathcal{O}] \simeq \sum_k w_k (\mathcal{O}(\boldsymbol{a}_k) - \mathrm{E}[\mathcal{O}])^2$$

$\rightarrow$ $\sum_k w_k = 1$

$\rightarrow$ unweighted sampling $w_1 = w_2 = \ldots$

$\rightarrow$ weighted sampling $w_1 \neq w_2 \neq \ldots$

- Here $\{w_k, \boldsymbol{a}_k\}$ is the sample distribution of the parent distribution $\mathcal{P}(\boldsymbol{a}|data)$
- Given the $\mathcal{P}(\boldsymbol{a}|data)$ the sample distribution is unique, regardless of the MC method

MC Method 1: data resampling

- Construct pseudo data sets where each data point is sampled using Gaussian distribution with mean and variance given by the original data

$$d_{k,i}^{(\text{pseudo})} = d_i^{(\text{exp})} + \sigma_i^{(\text{exp})} R_{k,i}$$

i : i–th data point

k : k–th pseudo data set index

$R_{k,i}$: random number from normal distribution

- Fit each pseudo data sample $k = 1, .., N$ to obtain parameter vectors $\boldsymbol{a}_k$ The sample distribution of $\mathcal{P}(\boldsymbol{a}|data)$ is approximately

$$\{w_k = 1/N, \boldsymbol{a}_k\}$$

here "fit" means Chi-square minimization

MC Method 1+: data resampling+cross validation

- **Issues with number of parameters**

→ Ideally one should not be worried about the number of parameters to be used.

→ This is an issue for Hessian method due to the flat directions.

→ However flat directions are typically only a local feature of the parent distribution.

- **Over-fitting**

→ If there are too many parameters there would be regions in the parameter space where $\mathcal{P}(\boldsymbol{a}|data)$ develops "spikes" → signal of over-fitting

→ One can use cross-validation to tame the "spikes"

MC Method 1+: data resampling+cross validation

- **Procedure**

→ For each pseudo data sample k split randomly the data set in $50/50$ and label them as "training" and "validation" respectively

→ Fit the "training" set and stop the fitting whenever the description of the "validation" set deteriorates → it avoids over-fitting

- **Caveat**

→ the resulting sample distribution is sensitive to the partition. Possible solutions include to rescale the uncertainties of the training and validation set to compensate for the splitting

MC Method 1+++: data resampling+cross validation +$a^{(\text{guess})}$ randomization +iterative runs

- **One vs. multiple minima**

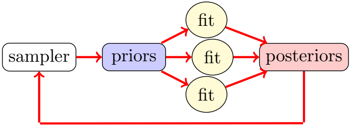
→ It is possible that $\mathcal{P}(\boldsymbol{a}|data)$ is multi modal.

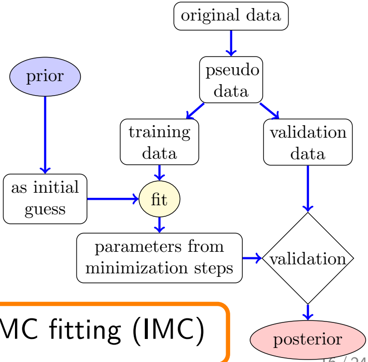
→ Hence it is important to start the scan from many different starting points

- **Caviat**

→ Optimization algorithms are based on gradient descent search. It is possible that in a given run with N independent scans the sample distribution does not represent accurately the "true" parent distribution

→ To solve this, we start a new run by sampling guessing parameters from the prior iteration





Iterative MC fitting (IMC)

MC Method 2: Hybrid Markov Chain Monte Carlo

- **The basic idea**

→ This is an MCMC based algorithm (random walks + rejection sampling)

→ The random walks are optimized by solving Hamilton's equations.

→ The parameters $\boldsymbol{a}$ are the "coordinates" and a conjugate vector $\boldsymbol{p}$ e.g. "momentum" is defined

→ An initial "state" is defined by a random coordinate vector $\boldsymbol{a}_0$ and a random momentum vector $\boldsymbol{p}_0$.

→ A new state is proposed by solving a Hamiltonian using the leap frog method

$$H(\boldsymbol{p}, \boldsymbol{a}) = \frac{\boldsymbol{p}^2}{2m} - \log(\mathcal{L}(\boldsymbol{a}))$$

- **pros**

→ It provides a faithful sampling distribution

- **cons**

→ the number of steps and step size of the leap frog must be tuned.

→ Cannot be parallelized

MC Method 3: nested resampling

- **The basic idea**: compute

$$Z = \int \mathcal{L}(\text{data}|\boldsymbol{a})\pi(\boldsymbol{a})d^n a = \int_0^1 \mathcal{L}(X)dX$$

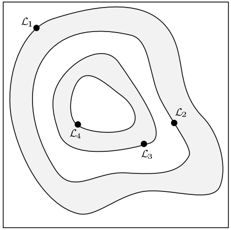
→ The algorithm traverses ordered isolikelihood contours in the variable X such that X follows the progression $X_i = t_i X_{i-1}$

→ The variable t_i is estimated statistically

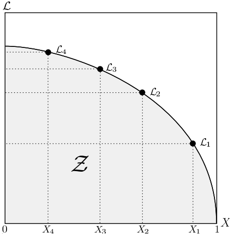
→ The algorithm can be optimized iteration to iteration. One can sample only in the regions where the likelihood is larger → "importance sampling"

→ The nested sampling is parallelizable

$\mathcal{L}(\text{data}|\boldsymbol{a})$ in $\boldsymbol{a}$ space

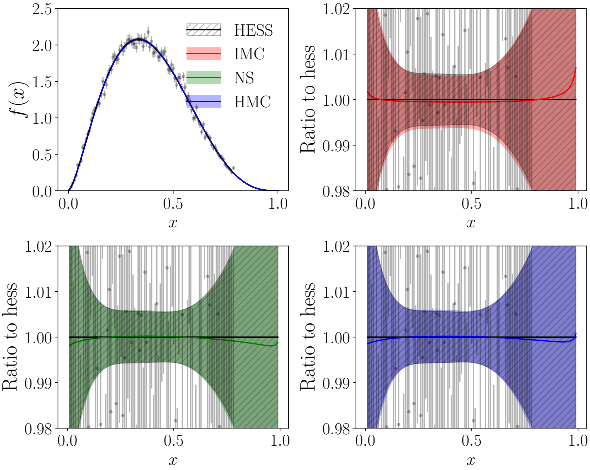


$\mathcal{L}(X)$ in X space



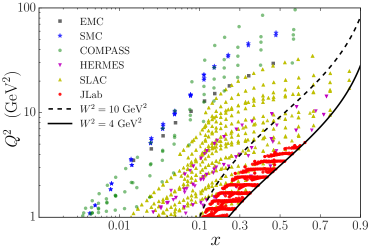
Toy example

- → We generate events from $f(x)$ to mimic realistic counting experiment
- → The fits and the error bands are performed with four different algorithms
- → Clearly all the methods give the same parent distribution for $f(x)$
- → This is expected as all the methods uses same likelihood

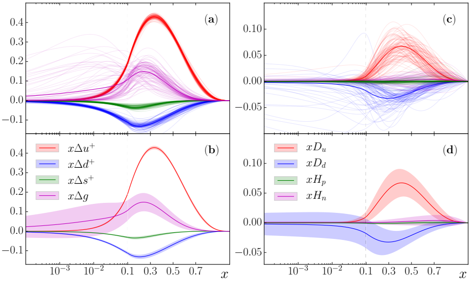


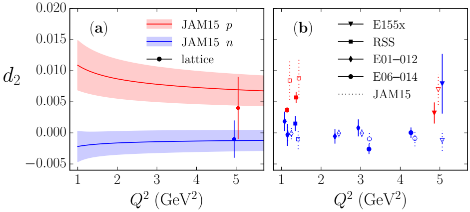
Polarized PDFs: inclusive polarized DIS

NS, Melnitchouk, Kuhn, Ethier, Accardi (PRD 93,074005)



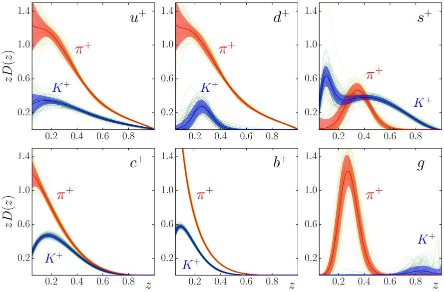
- $\rightarrow$ Inclusion of all the JLab 6GeV data
- $\rightarrow$ Determination of twist 3 g_2 (not power suppresed)
- $\rightarrow$ Extraction of d_2 matrix element

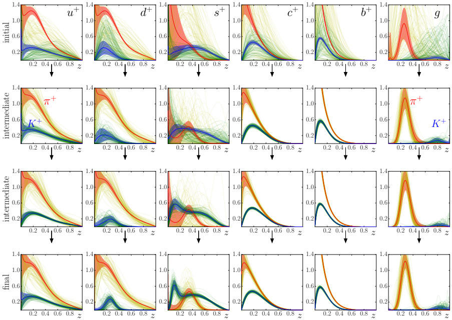




Fragmentation Functions: SIA

NS, Ethier, Melnitchouk, Hirai, Kumano, Accardi (PRD 94, 114004)





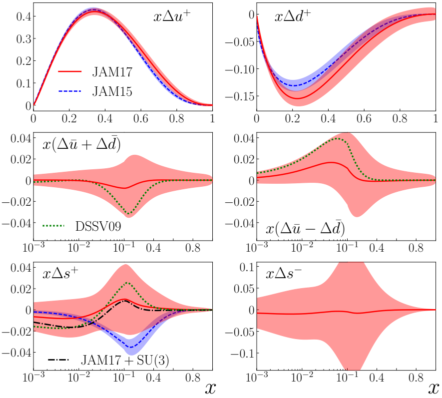
→ Inclusion of all the global data from Belle and Babar up to LEP data at $Q = M_z$

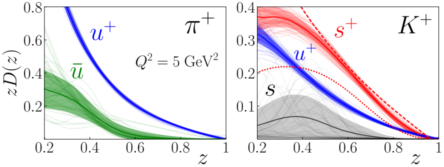
→ Fits were done for pion and kaon samples

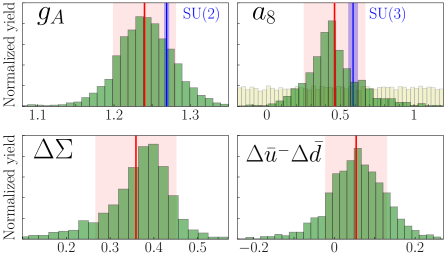
→ We only extracted $D_q^+ = D_q + D_{\bar{q}}$

Combined ΔPDF and FF: pDIS+pSIDIS+SIA

Ethier, NS, Melnitchouk (PRL 119, 132001)





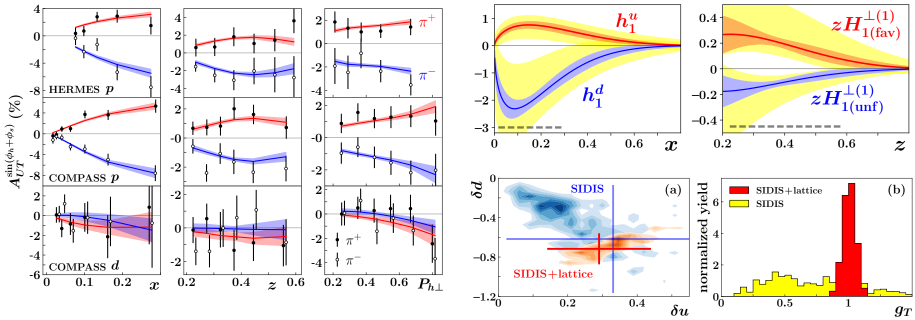


→ First simultaneous extraction of polarized PDFs and FFs

→ Extraction of the polarized strange distribution without SU(3) constraints

SIDIS+Lattice analysis of nucleon tensor charge

Lin, Melnitchouk, Prokudin, NS, Shows (arXiv:1710.09858)



→ Extraction of transversity and Collins FFs from SIDIS A_{UT}+Lattice g_T

→ In the absence of Lattice, SIDIS at present has no significant constraints on g_T → this will change with the upcoming JLab12 measurements

Summary and outlook

- → MC methods are becoming a very useful tool in QCD phenomenology.
- → It brings features that traditional methods cannot offer
- → Significant amount of research in data analysis is taking place outside of the field. Maybe it is time to modernize how we think and how we approach QCD global analyzes
- → In this talk I only covered "the tip of the iceberg", but there are many more interesting subtopics to be discussed e.g. treatment of incompatible data sets