Bayesian perspective on QCD global analysis

Nobuo Sato
University of Connecticut/JLab

DIS18,
Kobe, Japan,
April 16-20, 2018

In collaboration with:

- A. Accardi
- E. Nocera
- W. Melnitchouk

Jefferson Lab Angular Momentum Collaboration

Jefferson Lab
EXPLORING THE NATURE OF MATTER

Bayesian methodology in a nutshell

- In QCD global analysis PDFs are parametrized at some scale Q_0. e.g.

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

- "fitting" is essentially estimation of

$$\mathrm{E}[f] = \int d^n a \;\; \mathcal{P}(\boldsymbol{a}|data) \;\; f(\boldsymbol{a}) \qquad\qquad \boldsymbol{a} = (N,a,b,c,d,...)$$
$$\mathrm{V}[f] = \int d^n a \;\; \mathcal{P}(\boldsymbol{a}|data) \;\; (f(\boldsymbol{a}) - \mathrm{E}[f])^2$$

- The probability density $\mathcal{P}$ is given by the Bayes' theorem

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

Bayesian methodology in a nutshell

- The likelihood function is not unique. A standard choice is the Gaussian likelihood

$$\mathcal{L}(d|\boldsymbol{a}) = \exp\left[-\frac{1}{2}\sum_i\left(\frac{d_i - \mathrm{thy}_i(\boldsymbol{a})}{\delta d_i}\right)^2\right]$$

- Priors are design to veto unphysical regions in parameter space. e.g.

$$\pi(\boldsymbol{a}) = \prod_i \theta(a_i - a_i^{min})\theta(a_i^{max} - a_i)$$

- How do we compute $\mathrm{E}[f], \mathrm{V}[f]$?
  + Maximum likelihood
  + Monte Carlo methods

Maximum Likelihood

- Estimation of expectation value

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

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

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

- or equivalently Chi-squared minimization

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

Maximum Likelihood

- Estimation of variance (Hessian method)

$$\mathrm{V}[f] = \int d^n a \;\; \mathcal{P}(\boldsymbol{a}|data) \;\; (f(\boldsymbol{a}) - \mathrm{E}[f])^2$$
$$\simeq \sum_k \left(\frac{f(t_k = 1) - f(t_k = -1)}{2} \right)^2$$

- It relies on factorization of $\mathcal{P}(\boldsymbol{a}|data)$ along eigen directions

$$\mathcal{P}(\boldsymbol{a}|data) \propto \prod_k \exp\left(-\frac{1}{2} t_k^2\right) + O(\Delta a^3)$$

- and linear approximation of $f(\boldsymbol{a})$

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

Maximum Likelihood

- **pros**
  - + Very practical. Most PDF groups use this method
  - + It is computationally inexpensive
  - + f and its eigen directions can be precalculated/tabulated
- **cons**
  - + Assumes local Gaussian approximation of the likelihood
  - + Assumes linear approximation of the observables $\mathcal{O}$ around $\boldsymbol{a}_0$
  - + The assumptions are strictly valid for linear models.
  - + Computation of the Hessian matrix is numerically unstable if flat directions are present
- **examples**
  - → if $f(x) = a + bx + cx^2$ then $\mathrm{E}[f(x)] = \mathrm{E}[a] + \mathrm{E}[b]x + \mathrm{E}[c]x^2$
  - → 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

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

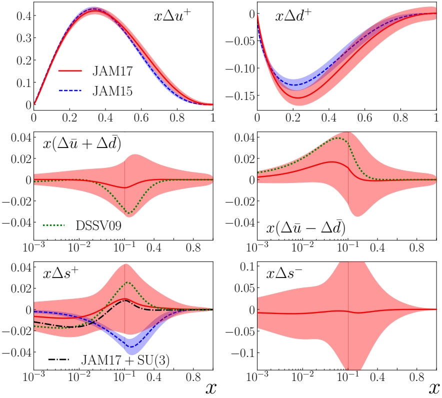
- Any MC method attempts to do this using MC sampling

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

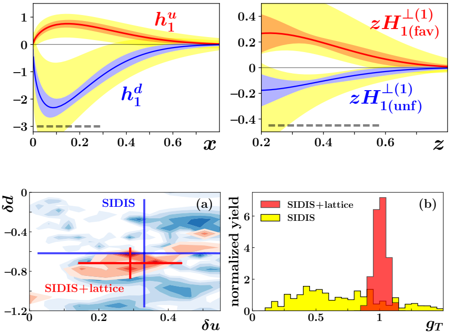
- i.e to construct the sample distribution $\{w_k, \boldsymbol{a}_k\}$ of the parent distribution $\mathcal{P}(\boldsymbol{a}|data)$

Monte Carlo Methods

- Resampling + cross validation
- Nested Sampling (NS)
- Hybrid Markov chain (HMC); Gabin Gbedo, Mangin-Brinet (2017)



resampling + CV, Ethier et al (2017)

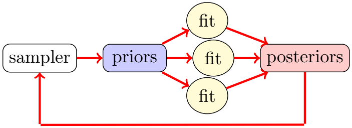


Nested Sampling, Lin et al (2018)

Resampling+cross validation (R+CV)

- Resample the data points within quoted uncertainties using Gaussian statistics

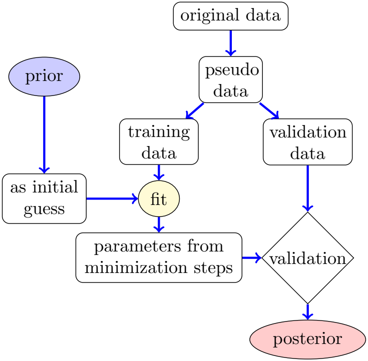
$$d_{k,i}^{(\text{pseudo})} = d_i^{(\text{exp})} + \sigma_i^{(\text{exp})} R_{k,i}$$



- Fit each pseudo data sample $k = 1, .., N$ to obtain parameter vectors $\boldsymbol{a}_k$:

$$\mathcal{P}(\boldsymbol{a}|data) \to \{w_k = 1/N, \boldsymbol{a}_k\}$$

- For large number of parameters, split the data into tranining and validation sets and find $\boldsymbol{a}_k$ that best describes the validation sample



Nested Sampling (NS)

- arXiv:astro-ph/0508461v2
- arXiv:astro-ph/0701867v2
- arxiv.org/abs/1703.09701

■ **The basic idea**: compute

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

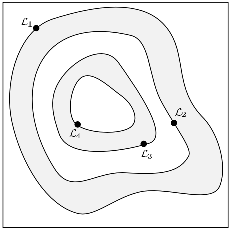
+ The procedure collects samples from isolikelihoods and they are weighted by their likelihood values

+ Insensitive to local minima $\rightarrow$ faithful conversion of

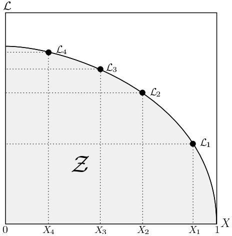
$$\mathcal{P}(\boldsymbol{a}|data) \rightarrow \{w_k, \boldsymbol{a}_k\}$$

+ Multiple runs can be combined into one single run $\rightarrow$ the procedure can be parallelized

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



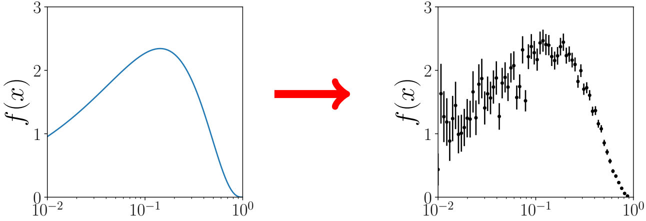
$\mathcal{L}(X)$ in X space



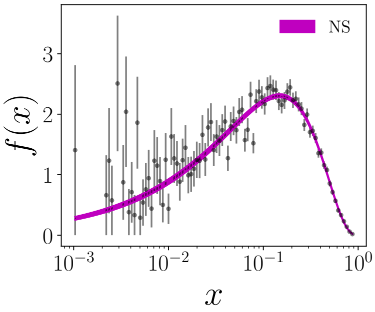
Comparison between the methods

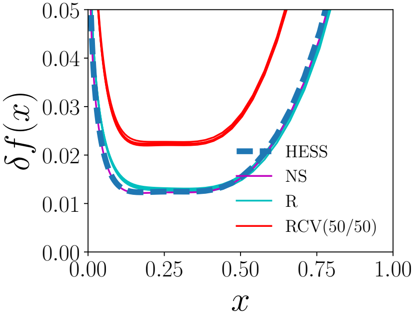
- Given a likelihood, does the evaluation of $\mathrm{E}[f]$ and $\mathrm{V}[f]$ depend on the method? $\rightarrow$ use stress testing numerical example

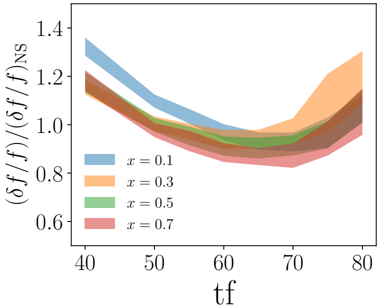
- Setup:
  - \+ Simulate a synthetic data via rejection sampling
  - \+ Estimate $\mathrm{E}[f]$ and $\mathrm{V}[f]$ using different methods



Comparison between the methods







- HESS, NS and R provide the same uncertainty
- R+CV over estimates the uncertainty by roughly a factor of 2
- Uncertainties also depends on training fraction (tf)
- The results confirmed also within a neural net parametrization

Beyond gaussian likelihood

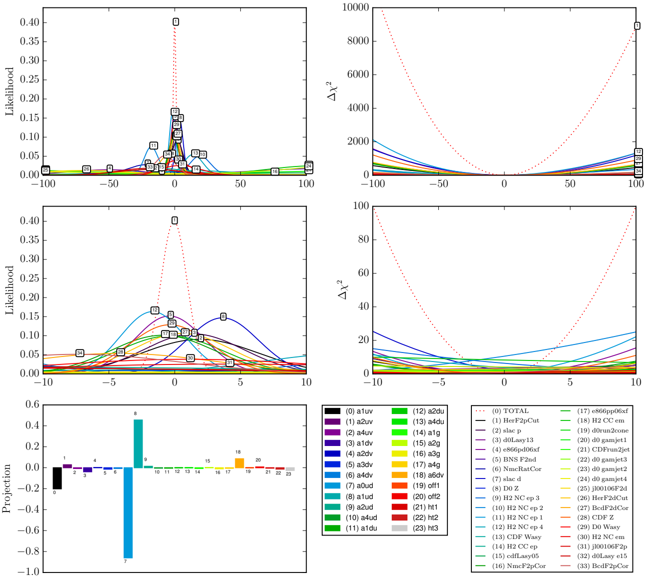
- The Gaussian likelihoods are not adequate to describe uncertainties in the presence of incompatible data sets
- Example:
  + Two measurements of a quantity m: $(m_1, \delta m_1)$, $(m_2, \delta m_2)$
  + The expectation value and variance can be computed exactly

$$\mathrm{E}[m] = \frac{m_1 \delta m_2 + m_2 \delta m_1}{\delta m_2^2 + \delta m_1^2}$$

$$\mathrm{V}[m] = \frac{\delta m_2^2 \delta m_1^2}{\delta m_2^2 + \delta m_1^2}$$

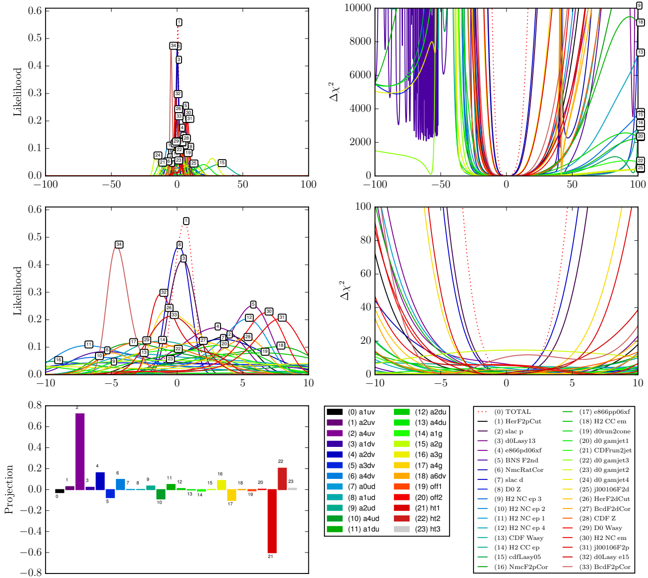
  + **note**: $\mathrm{V}[m]$ is independent of $|m_1 - m_2|$
- To obtain more realistic uncertainties, the likelihood function needs to be modified. (e.g. Tolerance criterion)

Likelihood profile in CJ15



- 24 parameters, 33 data sets
- Eigen direction without incompatibilities

Likelihood profile in CJ15



- 24 parameters, 33 data sets
- Eigen direction with incompatibilities
- Modified likelihood function is needed

Beyond gaussian likelihood

- Tolerance criterion (standard choice)
- Disjoint likelihood function. e.g.
  joint:

$$\mathcal{L}(m_1, m_2|m; \delta m_1 \delta m_2) = \mathcal{L}(m_1|m; \delta m_1)\mathcal{L}(m_2|m; \delta m_2)$$

$$\mathrm{E}[m] = \frac{m_1 \delta m_2 + m_2 \delta m_1}{\delta m_2^2 + \delta m_1^2} \qquad \mathrm{V}[m] = \frac{\delta m_2^2 \delta m_1^2}{\delta m_2^2 + \delta m_1^2}$$

  disjoint:

$$\mathcal{L}(m_1, m_2|m; \delta m_1 \delta m_2) = \frac{1}{2}(\mathcal{L}(m_1|m; \delta m_1) + \mathcal{L}(m_2|m; \delta m_2))$$

$$\mathrm{E}[m] = \frac{1}{2}(m_1 + m_2) \qquad \mathrm{V}[m] = \frac{1}{2}(\delta m_1^2 + \delta m_2^2) + \left(\frac{m_1 - m_2}{2}\right)^2$$

- Empirical Bayes, hierarchical Bayes ...
- Many alternatives still to be explored

Summary and outlook

+ Bayesian formulation for global analysis provides a more general perspective for global fits than the traditional chi-squared minimization

+ MC approaches are useful to explore new likelihood functions and priors

+ Uncertainties on PDFs depend on parametrization as well as assumptions about the likelihood function and the priors

+ Given the likelihood function and priors, uncertainties on PDFs should be independent of the parametrization in the region where PDFs can be constrained

+ Also the results should be independent of the MC sampling method