Bayesian perspective on global analysis of PDFs and FFs

Nobuo Sato
University of Connecticut/JLab

Seminar at Duke

Duke, 2018

The Bayesian framework for QCD global analysis

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

The Bayes theorem

- Consider a quantity f to be inferred from data
- The goal is to estimate $\mathcal{P}(f|data)$
- This is achieved by the **Bayes theorem**

$$\underbrace{\mathcal{P}(f|data)}_{\text{posterior}} = \underbrace{\frac{1}{Z}}_{\text{evidence}} \underbrace{\mathcal{L}(data|f)}_{\text{likelihood}} \underbrace{\pi(f)}_{\text{prior}}$$

Likelihoods and priors

- The **likelihood** function is typically chosen to be Gaussian

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

- The **prior** function allows to restrict forbidden values for f i.e.

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

- $\mathcal{P}(f|data)$ depends on what is chosen for $\mathcal{L}$ and π

Parametrization

- In practice f needs to be represented parametrized 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 Bayes theorem is implemented as

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

Expectation values and variances

- 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

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

Expectation values and variances

- 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
  $\rightarrow$ Not practical!
- The challenge: how to compute $\mathrm{E}[\mathcal{O}], \mathrm{V}[\mathcal{O}]$?
  - **Maximum likelihood**
  - **Monte Carlo approach**

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

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

- For Gaussian likelihood it is χ^2 minimization

$$\min\left[-2\log\left(\mathcal{L}(data|\boldsymbol{a})\pi(\boldsymbol{a})\right)\right] = -2\log\left(\mathcal{L}(data|\boldsymbol{a}_0)\pi(\boldsymbol{a}_0)\right)$$
$$= \chi^2(\boldsymbol{a}_0) - 2\log\left(\pi(\boldsymbol{a}_0)\right)$$

Hessian method : eigen direction decomposition

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

The posterior distribution "factorizes" along each eigen direction

Maximum Likelihood + Hessian method

- Estimation of variance

$$
\begin{aligned}
\mathrm{V}[\mathcal{O}] &= \int d^n a \;\; \mathcal{P}(\boldsymbol{a}|data) \;\; (\mathcal{O}(\boldsymbol{a}) - \mathrm{E}[\mathcal{O}])^2 \\
&\simeq \prod_k \int dt_k \frac{e^{-\frac{1}{2}t_k^2}}{\sqrt{2\pi}} \sum_{lm} \frac{\partial \mathcal{O}}{\partial t_l} \frac{\partial \mathcal{O}}{\partial t_m} t_l t_m \\
&= \sum_k \left(\frac{\partial \mathcal{O}}{\partial t_k} \right)^2 \simeq \sum_k \left[\frac{\mathcal{O}(t_k = 1) - \mathcal{O}(t_k = -1)}{2} \right]^2
\end{aligned}
$$

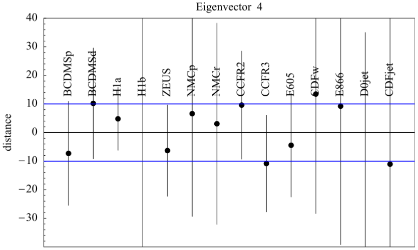
- It relies on
  - linear approximation for $\mathcal{O}(\boldsymbol{a})$
  - Gaussian factorization of the posterior

The tolerance criterion

- In QCD global analysis it is common to find discrepancies among datasets

- The variance is then scaled by a tolerance factor T

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



CTEQ6

$T \simeq 10$

Why do we need T?

Why do we need T?

- Consider observable m and measurements $(m_1, \delta m_1), (m_2, \delta m_2)$

- The χ^2 function is given by

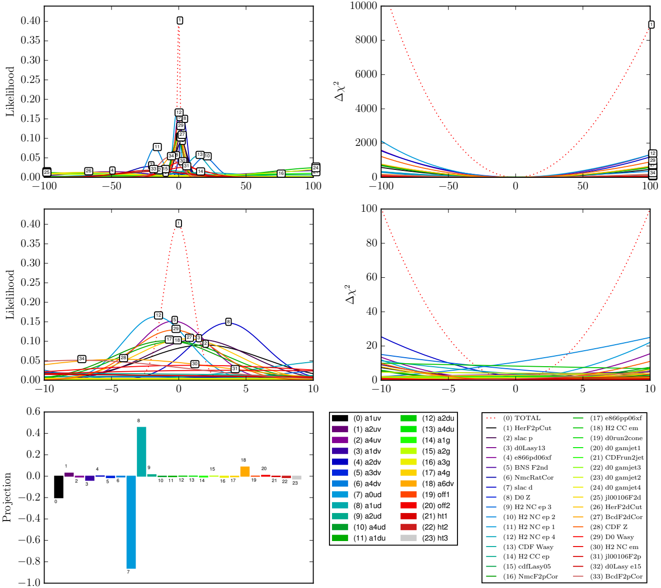
$$\chi^2(m) = \left(\frac{m - m_1}{\delta m_1}\right)^2 + \left(\frac{m - m_2}{\delta m_2}\right)^2$$

- The maximum likelihood and Hessian gives

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

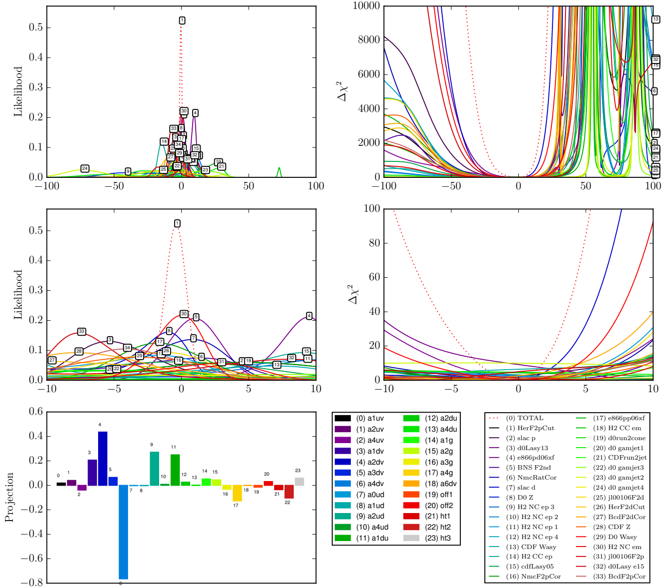
- $\mathrm{V}[m]$ is independent of $|m_1 - m_2|$

Real life global analysis of PDFs CJ15



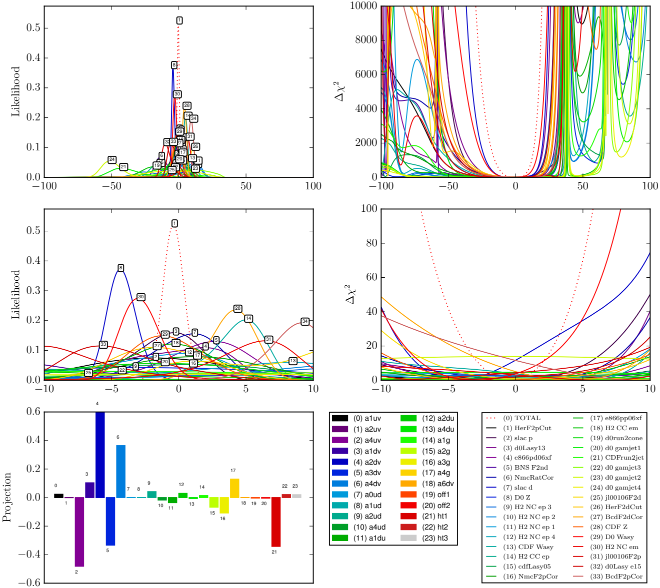
- eigen direction 1
- likelihood behaves as gaussian
- one can see which and datasets are relevant

Real life global analysis of PDFs CJ15



- eigen direction 13
- likelihood is less gaussian
- some datasets are in tension

Real life global analysis of PDFs CJ15



- eigen direction 16
- likelihood is less gaussian
- some datasets are in tension

Maximum Likelihood + Hessian method

- **pros**
  - Very practical. Most the 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 $\boldsymbol{a}_0$
  - These assumptions are strictly valid for linear models.
  - Hessian matrix is numerically unstable if flat directions are present
  - To deal with incompatible data one needs to apply the tolerance

Monte Carlo Methods

- Recall that we are interested in computing

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

- MC methods attempts to do this using MC sampling

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

- $\{w_k, \boldsymbol{a}_k\}$ is the **sample distribution** of the **posterior distribution** $\mathcal{P}(\boldsymbol{a}|data)$

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**

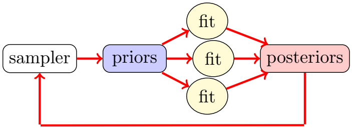
→ the resulting **sample distribution** is sensitive to the partition. Possible solutions include to rescale the uncertainties of the training and validation set to compensate for the splitting

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.

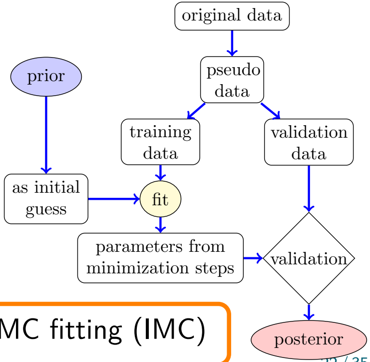
→ Hence it is important to start the scan from many different starting points



- **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

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

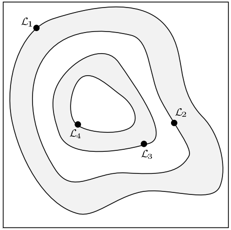
→ The algorithm traverses ordered isolikelihood contours in the variable X such that X follows the progression $X_i = t_i X_{i-1}$

→ The variable t_i is estimated statistically

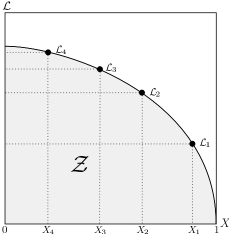
→ The algorithm can be optimized iteration to iteration. One can sample only in the regions where the likelihood is larger → "importance sampling"

→ The nested sampling is parallelizable

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



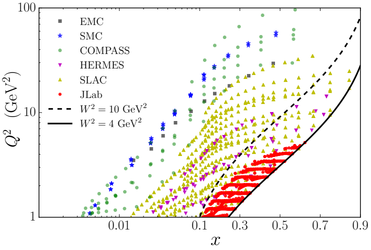
$\mathcal{L}(X)$ in X space

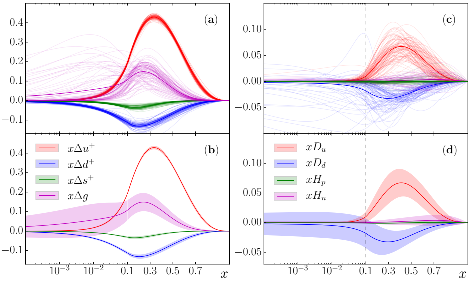


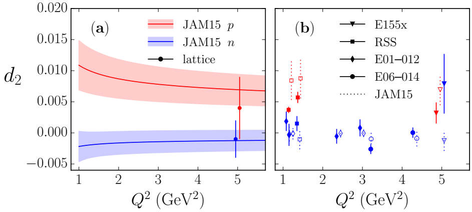
JAM global analysis

Polarized PDFs: inclusive polarized DIS

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



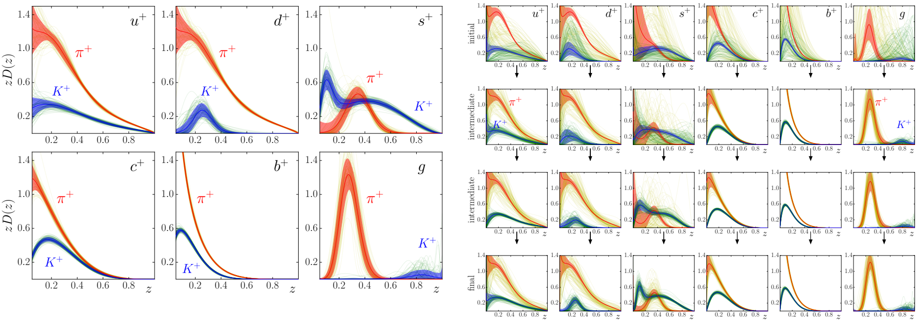




- → Inclusion of all the JLab 6GeV data
- → Determination of twist 3 g_2 (not power suppresed)
- → Extraction of d_2 matrix element

Fragmentation Functions: SIA

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



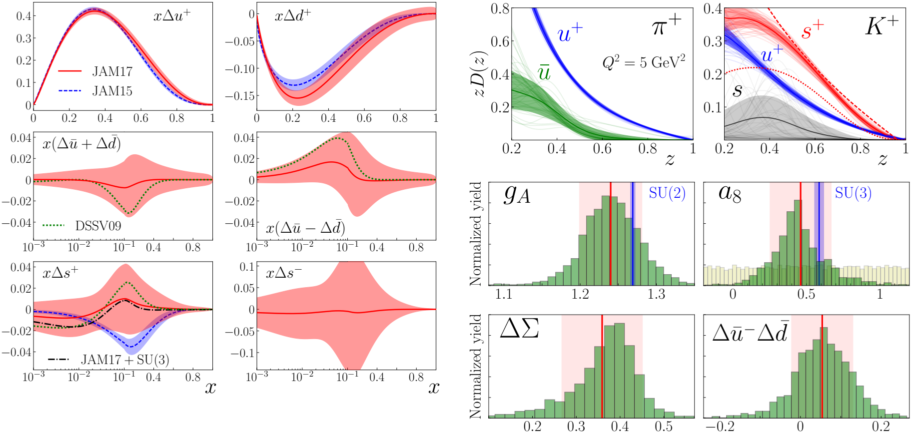
→ Inclusion of all the global data from Belle and Babar up to LEP data at $Q = M_z$

→ 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)

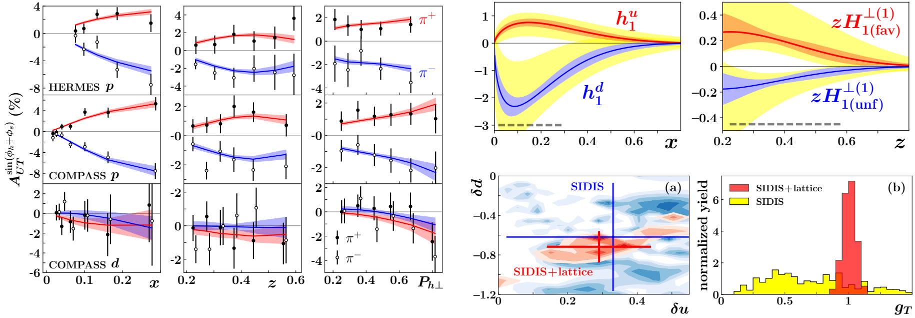


→ First simultaneous extraction of polarized PDFs and FFs

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

SIDIS+Lattice analysis of nucleon tensor charge

Lin, Melnitchouk, Prokudin, NS, Shows (PRLett 120, 152502)



$\rightarrow$ Extraction of transversity and Collins FFs from SIDIS A_{UT}+Lattice g_T

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

JAM Jefferson Lab Angular Momentum Collaboration

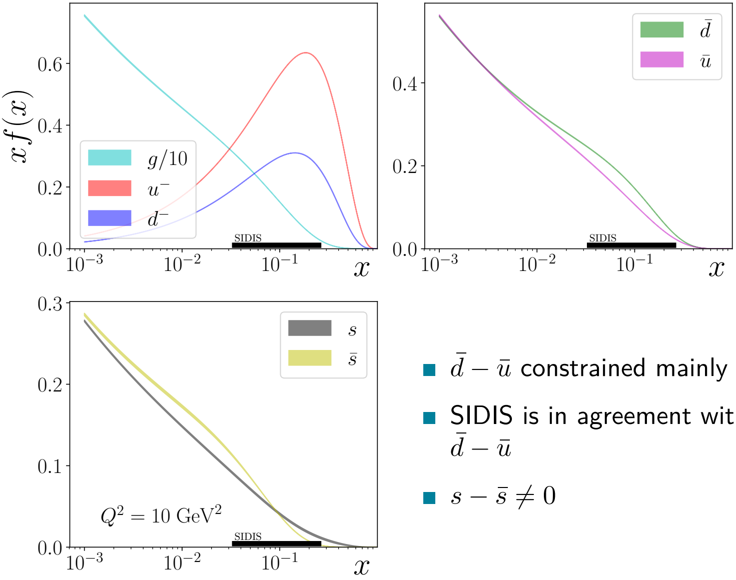
Present

JAM18: Universal analysis (preliminary)

Andres, Ethier, Melnitchouk, NS, Rogers

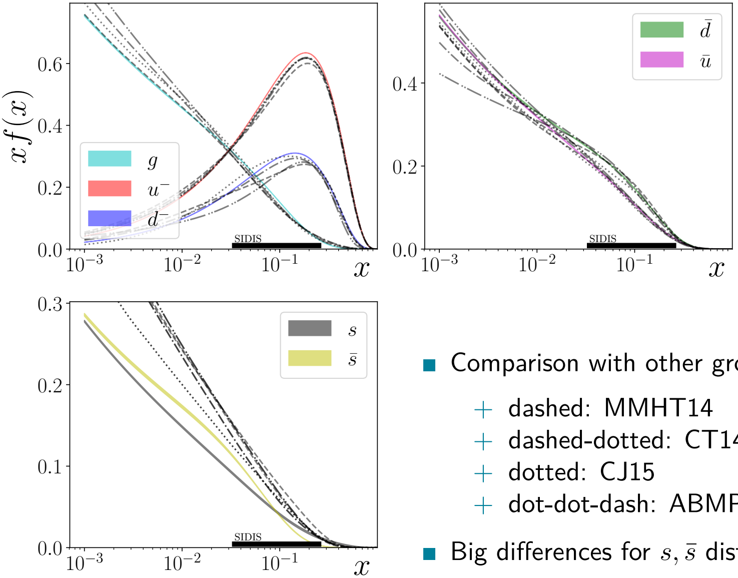
- **Data sets**
  - DIS, SIDIS(π, K), DY
  - ΔDIS, ΔSIDIS(π, K)
  - $e^+e^-(\pi, K)$
- **Theory setup**
  - Observables computed at **NLO in pQCD**
  - DIS structure functions only at **leading twist** $(W^2 > 10\ \mathrm{GeV}^2)$
- **Likelihood analysis (first steps)**
  - Use maximum likelihood to find a candidate solution
  - Use resampling to **check for stability** and estimate uncertainties
  - 80 shape parameters and 91 data normalization parameters: **171 dimensional space**
  - Sampling to be extended with IMC/Nested Sampling

JAM18: PDFs (preliminary)



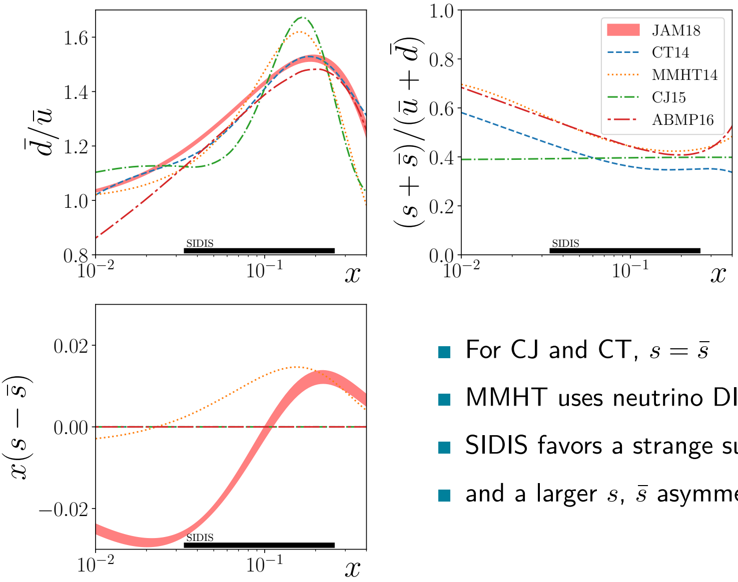
- $\bar{d} - \bar{u}$ constrained mainly by DY
- SIDIS is in agreement with DY's $\bar{d} - \bar{u}$
- $s - \bar{s} \neq 0$

JAM18: PDFs (preliminary)



- Comparison with other groups
  - dashed: MMHT14
  - dashed-dotted: CT14
  - dotted: CJ15
  - dot-dot-dash: ABMP16
- Big differences for $s, \bar{s}$ distributions

JAM18: upolarized sea (preliminary)



- For CJ and CT, $s = \bar{s}$
- MMHT uses neutrino DIS
- SIDIS favors a strange suppression
- and a larger s, $\bar{s}$ asymmetry

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