HUGS Topical Seminar
June 16, 2021

PDF fitting: methods and uncertainties

Wally Melnitchouk

Jefferson Lab

Parton distributions in hadrons

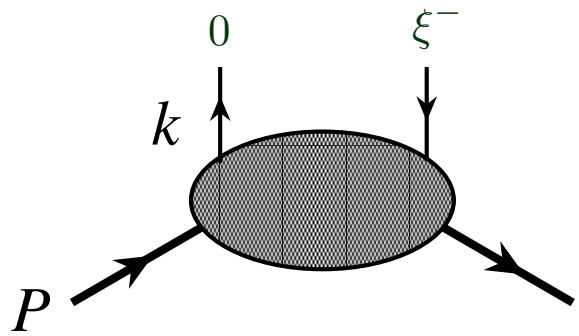
- Parton distribution functions (PDFs) are light-cone correlation functions

$$q(x) = \int_{-\infty}^{\infty} d\xi^- e^{-ixP^+\xi^-} \langle P | \overline{\psi}(\xi^-) \gamma^+ \mathcal{W}(\xi^-, 0) \psi(0) | P \rangle$$

→ light cone momentum fraction $x = \dfrac{k^+}{P^+}$

→ Wilson line (gauge invariance)

$$\mathcal{W}(\xi^-, 0) = \exp\left\{ - ig \int_0^{\xi^-} d\eta^- \mathcal{A}^+(\eta^-) \right\}$$



- In $\mathcal{A}^+ = 0$ gauge, in fast-moving frame PDF has a probabilistic interpretation as a particle density

$$\int_{-1}^{1} dx\, q(x) = \langle P | \overline{\psi}(0) \gamma^+ \psi(0) | P \rangle \approx \langle P | \psi^\dagger(0) \psi(0) | P \rangle$$

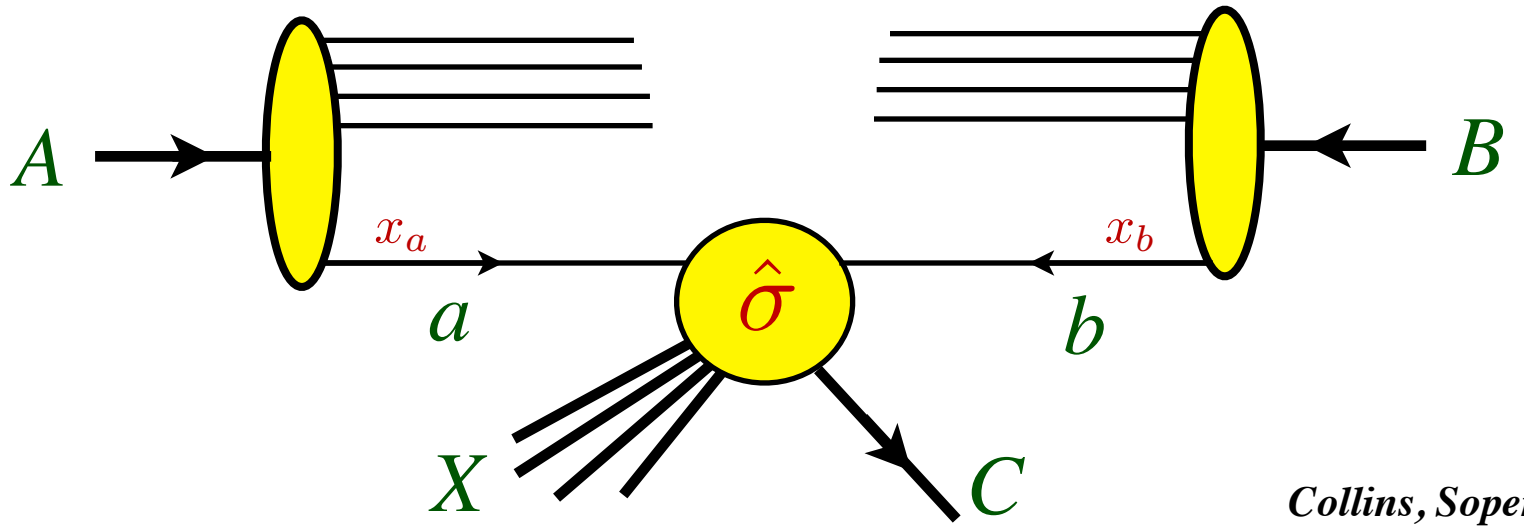
number density

$\overline{\psi} \gamma^0 \psi \approx \overline{\psi} \gamma^z \psi$

number operator

Parton distributions in hadrons

- Inclusive high-energy particle production $A\,B \to C\,X$



Collins, Soper, Sterman (1980s)

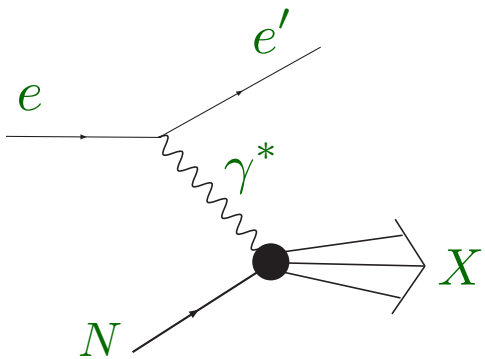
→ QCD factorization: separation of hard (perturbative, calculable) from soft (nonperturbative, parametrized) physics

$$\sigma_{AB\to CX}(p_A,p_B) = \sum_{a,b} \int dx_a\, dx_b\, \underbrace{f_{a/A}(x_a,\mu)}_{\cdots\cdots}\, \underbrace{f_{b/B}(x_b,\mu)}_{\cdots\cdots} \times \sum_n \alpha_s^n(\mu)\, \hat{\sigma}^{(n)}_{ab\to CX}\,(x_a p_A, x_b p_B, Q/\mu)$$

→ process-independent parton distribution functions $f_{a/A}$ characterizing structure of bound state A

Parton distributions in hadrons

- Most information on PDFs obtained from lepton-hadron deep-inelastic scattering (DIS)



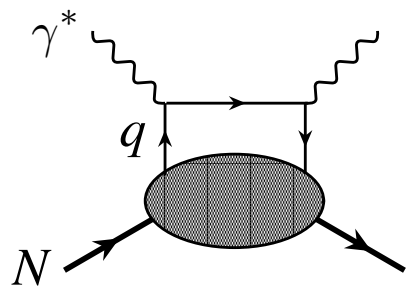
$$\frac{d^2\sigma}{d\Omega dE'} = \frac{4\alpha^2 E'^2 \cos^2\frac{\theta}{2}}{Q^4}\left(2\tan^2\frac{\theta}{2}\frac{F_1}{M} + \frac{F_2}{\nu}\right)$$

$$x_B = \frac{Q^2}{2M\nu} \qquad Q^2 = \vec{q}^{\,2} - \nu^2 \qquad \nu = E - E'$$

→ structure function given as convolution of hard Wilson coefficient with PDF

$$F_2(x_B, Q^2) = x_B \sum_q e_q^2 \int_{x_B}^1 \frac{dx}{x} C_q\left(\frac{x_B}{x}, \alpha_s\right) q(x, Q^2)$$

$$\rightarrow x_B \sum_q e_q^2\, q(x_B, Q^2)$$

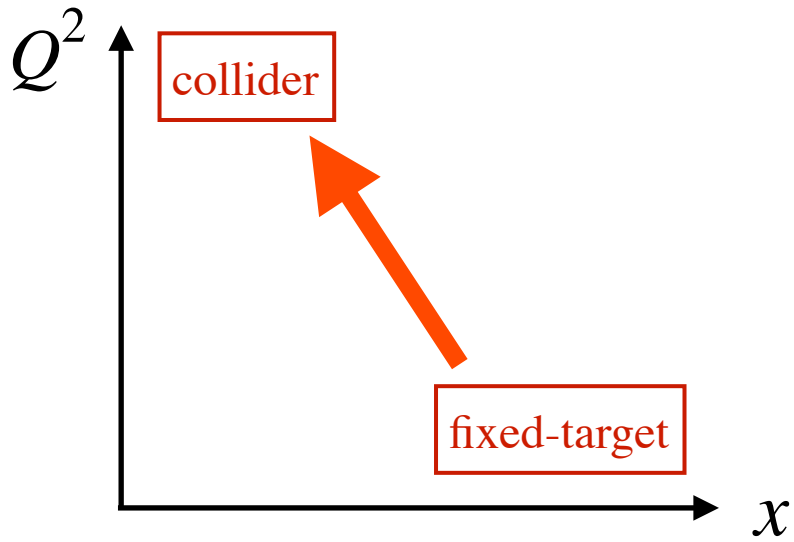


for leading order approximation $C_q \rightarrow \delta\left(1 - \frac{x_B}{x}\right)$

Parton distributions in hadrons

- Precision PDFs needed to
  (1) understand basic structure of QCD bound states
  (2) compute backgrounds in searches for BSM physics

→ Q^2 evolution feeds
low x, high Q^2 ("LHC")
from high x, low Q^2 ("JLab")



- Information on PDFs obtained from
  (1) nonperturbative approaches (low-energy models, DSE, χEFT)
  (2) lattice QCD
  (3) global QCD analysis

Global PDF analysis

- Universality of PDFs allows data from many different processes (DIS, SIDIS, weak boson/jet production in pp, Drell-Yan …) to be analyzed simultaneously

  → distributions parametrized using a specific functional form, with parameters fitted to data

  $$x f(x,\mu) = N x^{\alpha} (1-x)^{\beta} P(x)$$

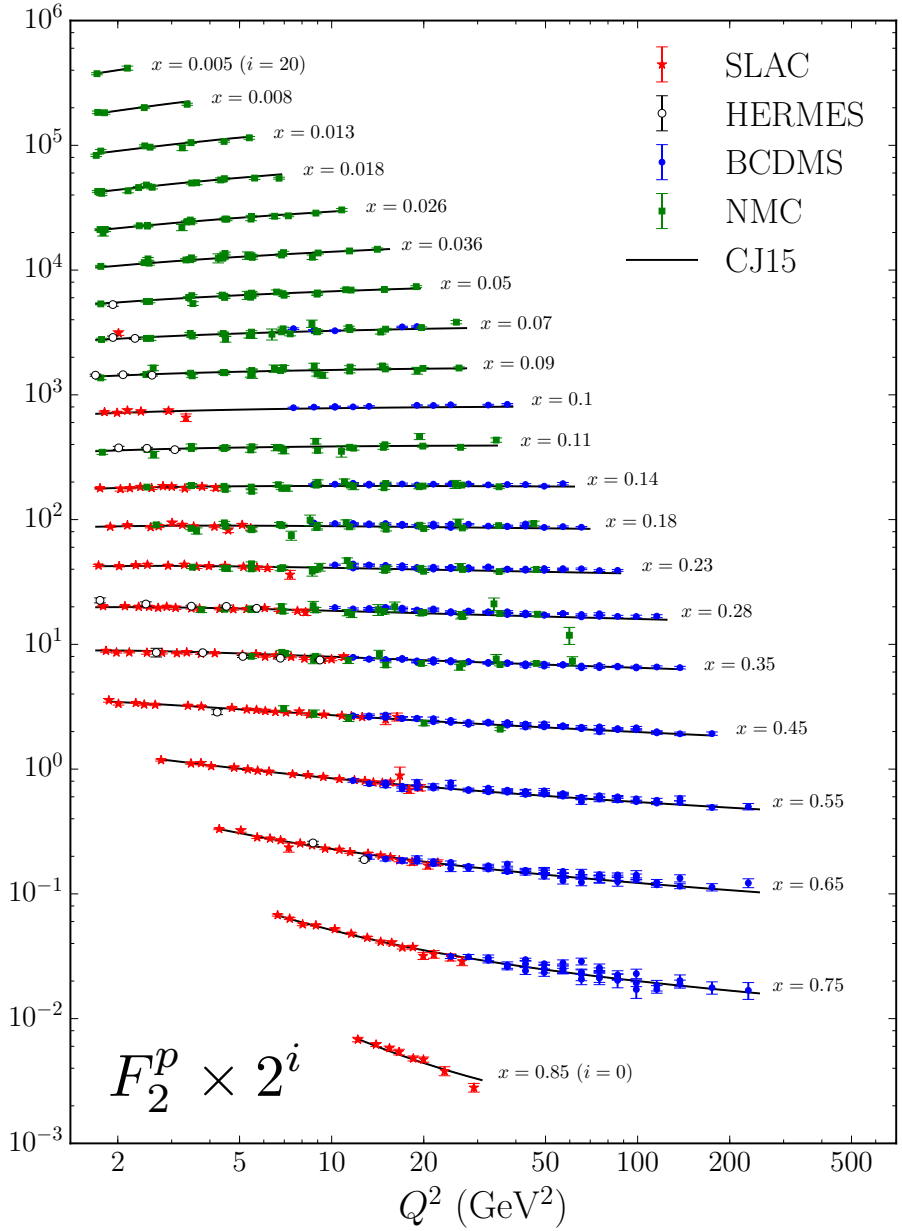
  with polynomial *e.g.* $P(x) = 1 + \epsilon\sqrt{x} + \eta\, x$
  or Chebyshev, neural net, …

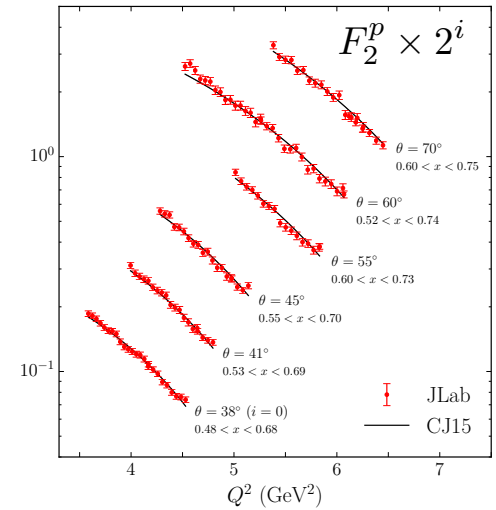
- Extraction of PDFs is challenging because usually there exist multiple solutions — "inverse problem"

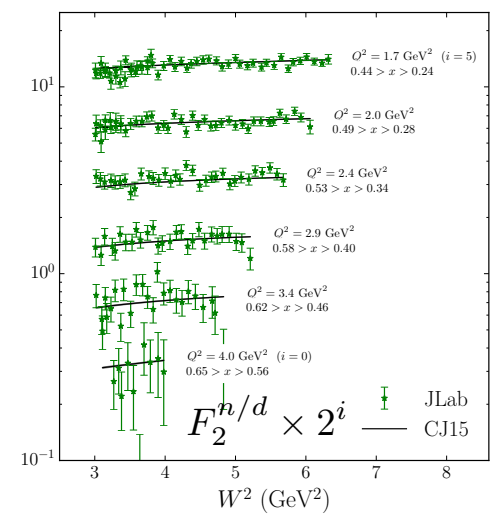
  → PDFs are not directly measured, but inferred from observables involving convolutions with other functions

Global PDF analysis

- Modern global QCD analyses typically fit 1000s of data points from high-energy scattering experiments

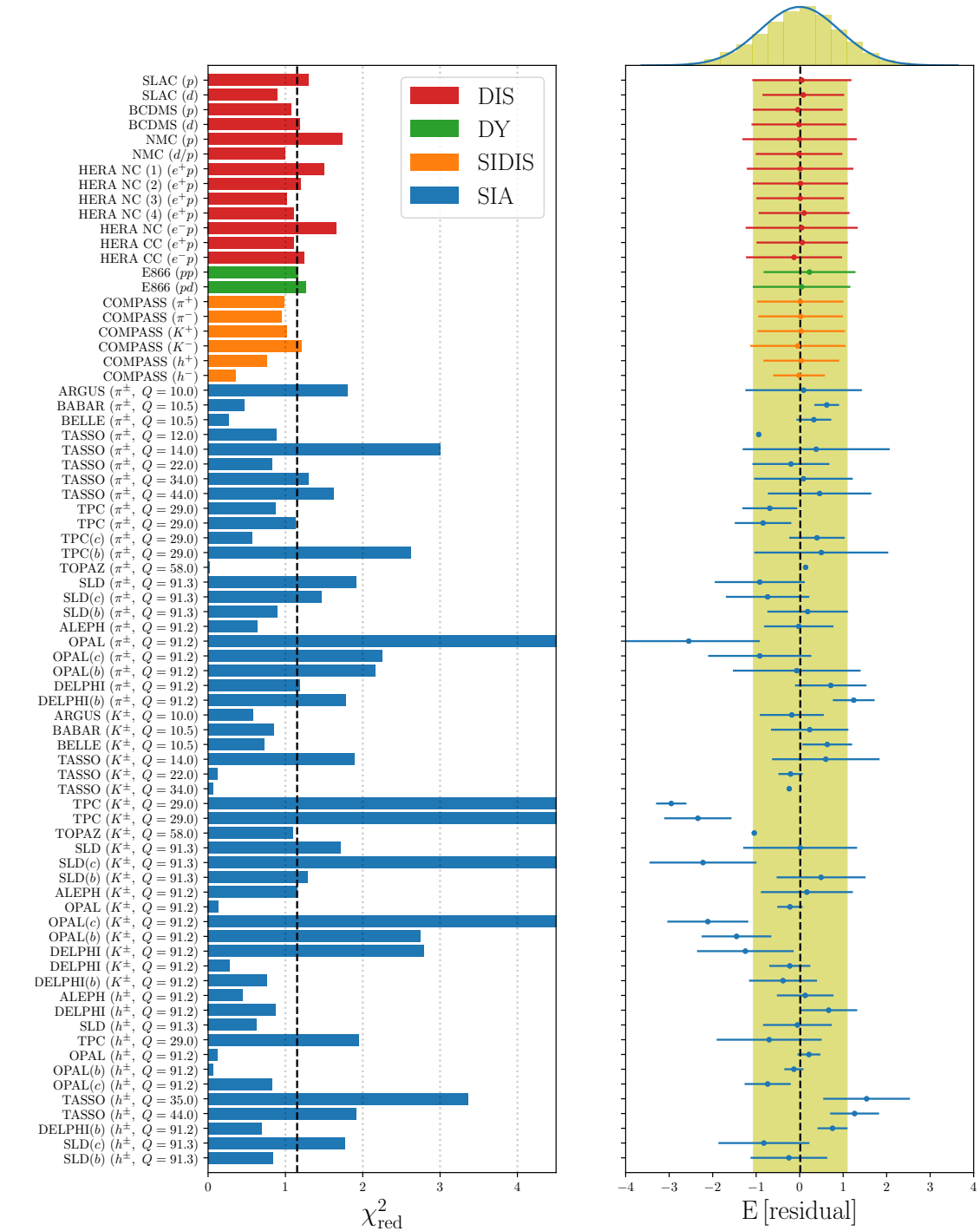






Accardi et al. ("CJ15")
PRD93, 114017 (2016)

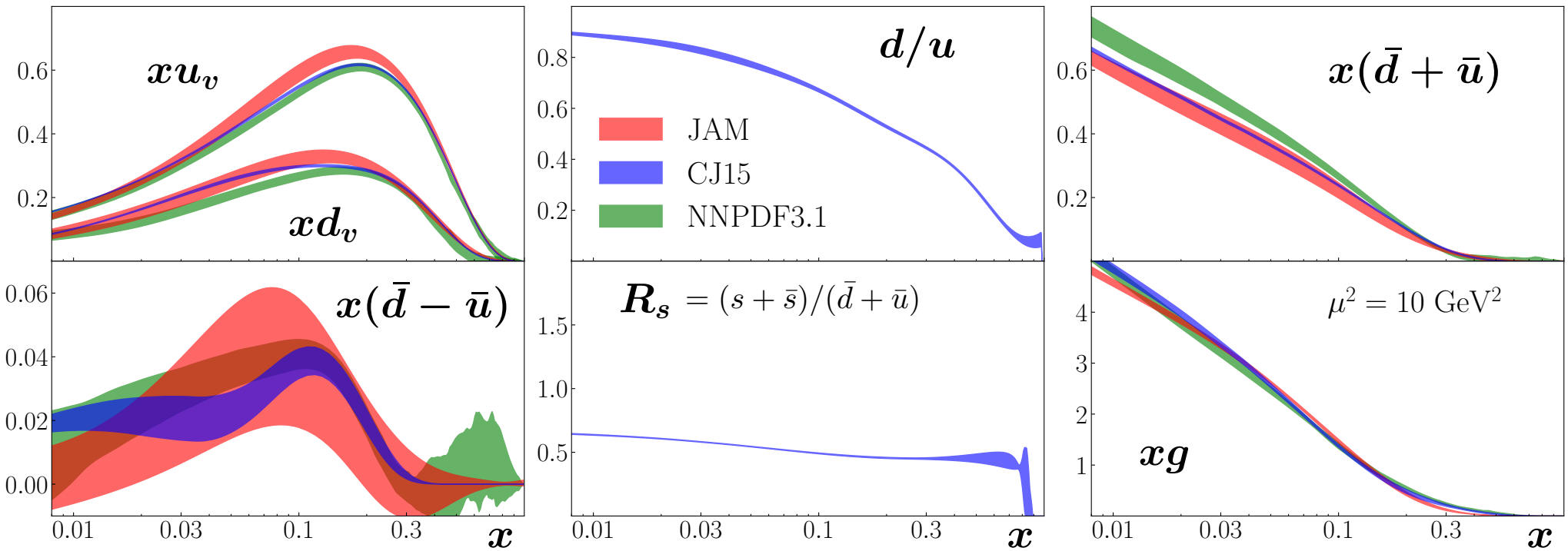
Global PDF analysis



Moffat, WM, Rogers, Sato
arXiv:2101.04664 ("JAM20")

Global PDF analysis

JAM20-SIDIS



Moffat, WM, Rogers, Sato
arXiv:2101.04664 ("JAM20")

Why are PDF uncertainties important?

- In searches for new physics beyond the Standard Model a major source of uncertainty on limits/discoveries is from calculation of QCD backgrounds $\longrightarrow$ PDF errors!

> "The PDF and α_s uncertainties were calculated using the PDF4LHC prescription [39] with the MSTW2008 68% CL NNLO [40, 41], CT10 NNLO [42, 43], and NNPDF2.3 5f FFN [44] PDF sets, and added in quadrature to the scale uncertainty."
>
> *Measurements of the charge asymmetry in top-quark pair production in the dilepton final state at $\sqrt{s}$ = 8 TeV with the ATLAS detector* PRD **94**, 032006 (2016)

$\longrightarrow$ drives a large part of the global PDF community (esp. LHC)

- Limits understanding of nucleon structure

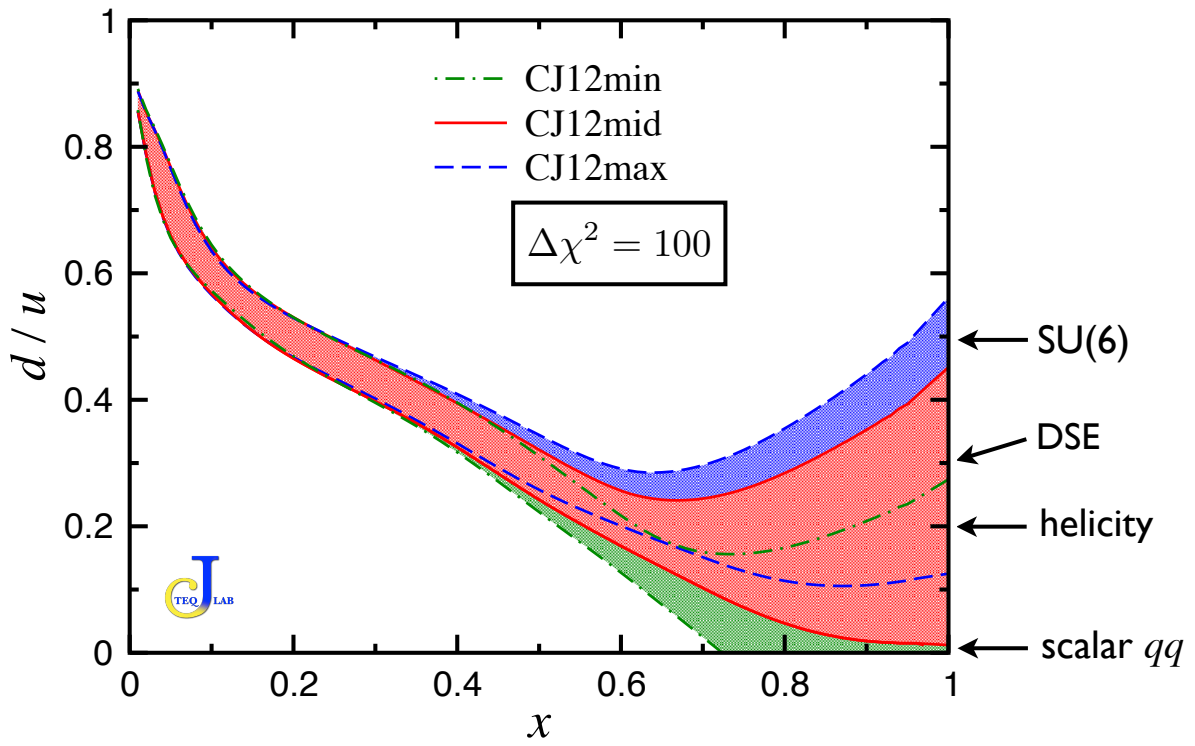
$\longrightarrow$ *e.g.* momentum and spin distributions of d quarks at large x

$\longrightarrow$ motivation for several JLab12 experiments
(MARATHON, BONuS, SoLID, …)

Why are PDF uncertainties important?

- Traditionally extracted from neutron / proton structure function ratio (where "neutron" ~ deuteron − proton), but large nuclear uncertainties affect high-x region

→ cannot discriminate between predictions for d/u at $x \sim 1$

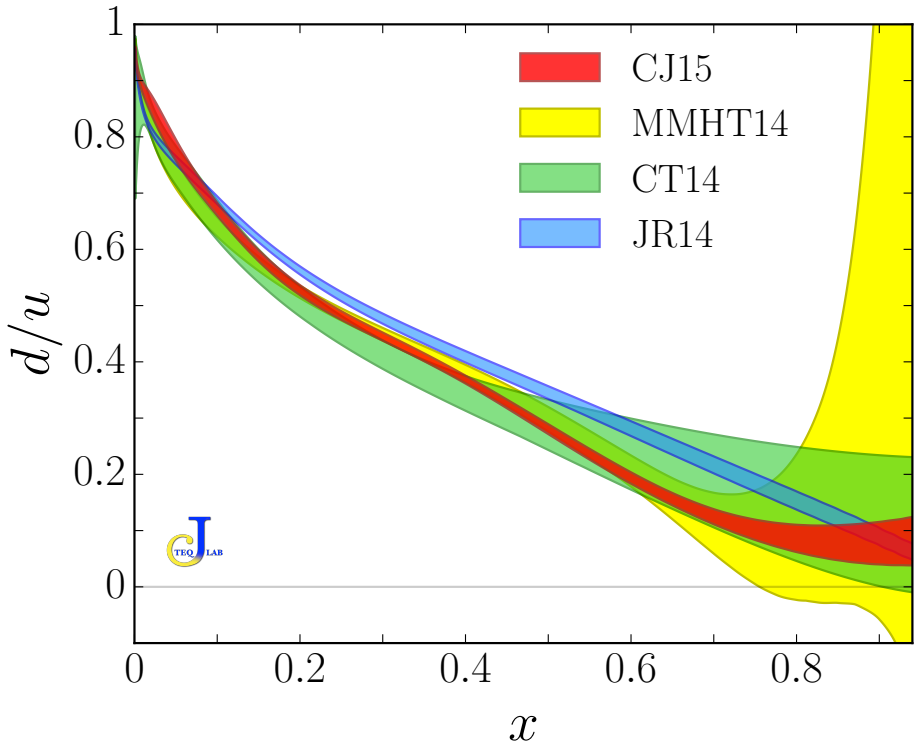


Owens, Accardi, WM (2013)

Why are PDF uncertainties important?

- Different groups use different definitions of PDF uncertainties to take into account *tensions* between data sets

→ multiply uncertainties by "tolerance" factor $T = \sqrt{\Delta\chi^2}$



→ CJ15: $\Delta\chi^2 = 2.7$

→ MMHT: $\Delta\chi^2 \approx 25 - 100$

→ CT14: $\Delta\chi^2 \approx 100$

→ JR14: $\Delta\chi^2 = 1$

… is this a meaningful comparison?

Need for new technology

- A major challenge has been to characterize PDF uncertainties — in a statistically meaningful way — in the presence of *tensions* among data sets

- Previous attempts sought to address tensions in data sets by introducing

  → "tolerance" factors (artificially inflating PDF errors)

  → "neural net" parametrization (instead of polynomial parametrization), together with MC techniques

- However, to address the problem in a more statistically rigorous way, one requires going *beyond* the standard χ^2 minimization paradigm

  → utilize modern techniques based on Bayesian statistics!

Bayesian approach to fitting

- Analysis of data requires estimating expectation values E and variances V of "observables" $\mathcal{O}$ (= PDFs, FFs) which are functions of parameters $\vec{a}$

$$E[\mathcal{O}] = \int d^n a \, \mathcal{P}(\vec{a}|\text{data}) \, \mathcal{O}(\vec{a})$$

$$V[\mathcal{O}] = \int d^n a \, \mathcal{P}(\vec{a}|\text{data}) \left[\mathcal{O}(\vec{a}) - E[\mathcal{O}]\right]^2$$

"Bayesian master formulas"

- Using Bayes' theorem, probability distribution $\mathcal{P}$ given by

$$\mathcal{P}(\vec{a}|\text{data}) = \frac{1}{Z} \mathcal{L}(\text{data}|\vec{a}) \, \pi(\vec{a})$$

in terms of the likelihood function $\mathcal{L}$

Bayesian approach to fitting

- Likelihood function

$$\mathcal{L}(\text{data}|\vec{a}) = \exp\left(-\frac{1}{2}\chi^2(\vec{a})\right)$$

is a Gaussian form in the data, with χ^2 function

$$\chi^2(\vec{a}) = \sum_i \left(\frac{\text{data}_i - \text{theory}_i(\vec{a})}{\delta(\text{data})}\right)^2$$

with priors $\pi(\vec{a})$ and "evidence" Z

$$Z = \int d^n a\, \mathcal{L}(\text{data}|\vec{a})\, \pi(\vec{a})$$

$\longrightarrow$ Z tests if *e.g.* an n-parameter fit is statistically different from $(n{+}1)$-parameter fit

Bayesian approach to fitting

- Two methods generally used for computing Bayesian master formulas:

Maximum Likelihood
(χ^2 minimization)

Monte Carlo

Bayesian approach to fitting

- Two methods generally used for computing Bayesian master formulas:

Maximum Likelihood
(χ^2 minimization)

⟶ maximize probability distribution $\mathcal{P}$ by minimizing χ^2 for a set of best-fit parameters $\vec{a}_0$

$$E\left[\vec{a}\right] = \vec{a}_0$$

⟶ if $\mathcal{O}$ is $\approx$ linear in the parameters, and if probability is symmetric in all parameters

$$E\left[\mathcal{O}(\vec{a})\right] \approx \mathcal{O}(\vec{a}_0)$$

Bayesian approach to fitting

- Two methods generally used for computing Bayesian master formulas:

Maximum Likelihood
(χ^2 minimization)

$\longrightarrow$ variance computed by expanding $\mathcal{O}(\vec{a})$ about $\vec{a}_0$
$e.g.$ in 1 dimension have "master formula"

$$V[\mathcal{O}] \approx \frac{1}{4}\Big[\mathcal{O}(a+\delta a) - \mathcal{O}(a-\delta a)\Big]^2$$

where

$$\delta a^2 = V[a]$$

Bayesian approach to fitting

- Two methods generally used for computing Bayesian master formulas:

Maximum Likelihood
(χ^2 minimization)

$\longrightarrow$ generalization to multiple dimensions via Hessian approach:

find set of (orthogonal) contours in parameter space around $\vec{a}_0$ such that $\mathcal{L}$ along each contour is parametrized by statistically independent parameters — directions of contours given by eigenvectors $\hat{e}_k$ of Hessian matrix H, with elements

$$H_{ij} = \frac{1}{2} \left. \frac{\partial^2 \chi^2(\vec{a})}{\partial a_i \partial a_j} \right|_{\vec{a}=\vec{a}_0}$$

and contours parametrized as $\Delta a^{(k)} = a^{(k)} - a_0 = t_k \dfrac{\hat{e}_k}{\sqrt{v_k}}$, with v_k eigenvectors of H

Bayesian approach to fitting

- Two methods generally used for computing Bayesian master formulas:

Maximum Likelihood
(χ^2 minimization)

→ basic assumption: $\mathcal{P}$ factorizes along each eigendirection

$$\mathcal{P}(\Delta a) \approx \prod_k \mathcal{P}_k(t_k)$$

where

$$\mathcal{P}_k(t_k) = \mathcal{N}_k \exp\Big[-\frac{1}{2}\chi^2\Big(a_0 + t_k \frac{\hat{e}_k}{\sqrt{v_k}}\Big)\Big]$$

note: in quadratic approximation for χ^2, this becomes a normal distribution

Bayesian approach to fitting

- Two methods generally used for computing Bayesian master formulas:

Maximum Likelihood
(χ^2 minimization)

→ uncertainties on $\mathcal{O}$ along each eigendirection
(assuming linear approximation)

$$(\Delta\mathcal{O}_k)^2 \approx \frac{1}{4}\left[\mathcal{O}\left(a_0 + T_k\frac{\hat{e}_k}{\sqrt{v_k}}\right) - \mathcal{O}\left(a_0 - T_k\frac{\hat{e}_k}{\sqrt{v_k}}\right)\right]^2$$

where T_k is finite step size in t_k, with total variance

$$V[\mathcal{O}] = \sum_k (\Delta\mathcal{O}_k)^2$$

Bayesian approach to fitting

- Two methods generally used for computing Bayesian master formulas:

Monte Carlo

→ in practice, generally one has $E[\mathcal{O}(\vec{a})] \neq \mathcal{O}(E[\vec{a}])$
so the maximal likelihood method will sometimes fail

→ Monte Carlo approach samples parameter space and assigns weights w_k to each set of parameters a_k

→ expectation value and variance are then weighted averages

$$E[\mathcal{O}(\vec{a})] = \sum_k w_k \, \mathcal{O}(\vec{a}_k), \qquad V[\mathcal{O}(\vec{a})] = \sum_k w_k \left(\mathcal{O}(\vec{a}_k) - E[\mathcal{O}]\right)^2$$

Bayesian approach to fitting

- Two methods generally used for computing Bayesian master formulas:

Maximum Likelihood
(χ^2 minimization)

- fast
- assumes Gaussianity
- no guarantee that global minimum has been found
- errors only characterize local geometry of χ^2 function

Monte Carlo

- slow
- does not rely on Gaussian assumptions
- includes all possible solutions
- accurate

Incompatible data sets

- Incompatible data sets can arise because of errors in determining central values, or underestimation of systematic experimental uncertainties

  $\longrightarrow$ requires some sort of modification to standard statistics

- Modify the master formula by introducing a "tolerance" factor T

$$V[\mathcal{O}] \;\rightarrow\; T^2\, V[\mathcal{O}]$$

  $e.g.$ for one dimension

$$V[\mathcal{O}] = \frac{T^2}{4}\Big[\mathcal{O}(a+\delta a) - \mathcal{O}(a-\delta a)\Big]^2$$

  $\longrightarrow$ effectively modifies the likelihood function

Incompatible data sets

- Simple example: consider observable m, and two measurements

$$(m_1, \delta m_1), \quad (m_2, \delta m_2)$$

$\longrightarrow$ compute exactly the χ^2 function

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

and, from Bayesian master formula, the mean value

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

and variance

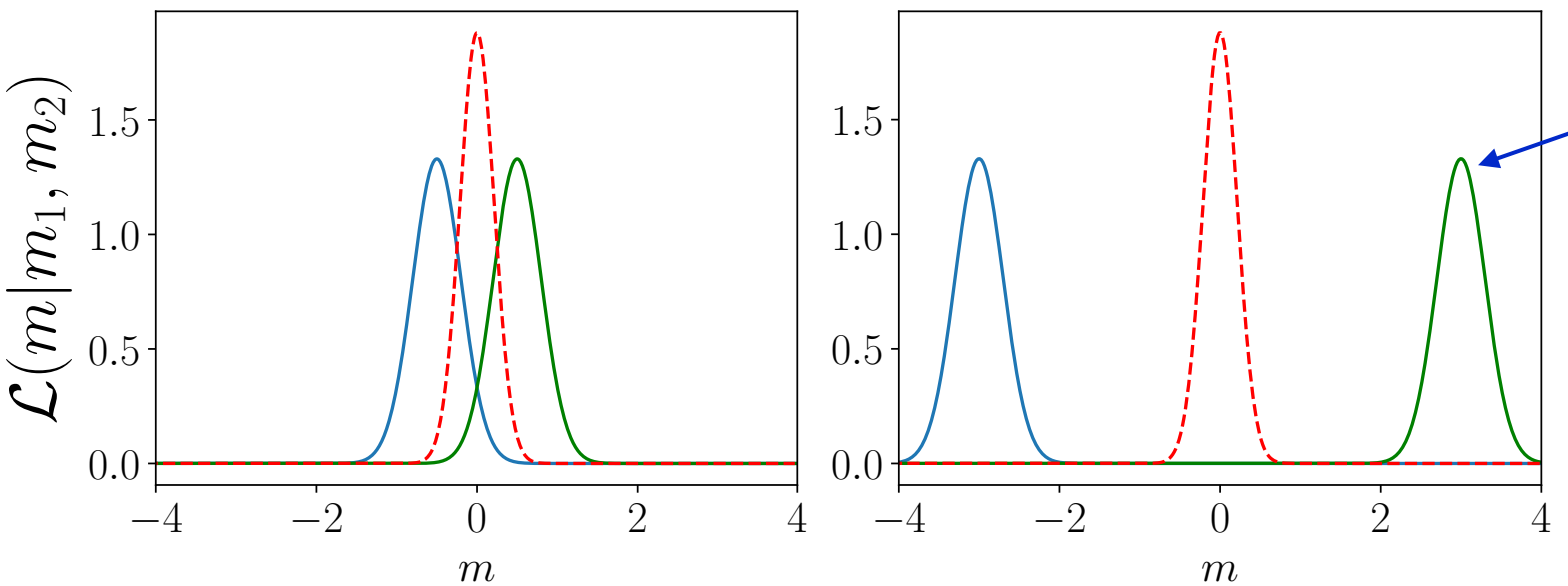
$$V[m] = H^{-1} = \frac{\delta m_1^2 \, \delta m_2^2}{\delta m_1^2 + \delta m_2^2}$$

does not depend on $m_1 - m_2$!

Incompatible data sets

- Simple example: consider observable m, and two measurements

$$(m_1, \delta m_1), \quad (m_2, \delta m_2)$$



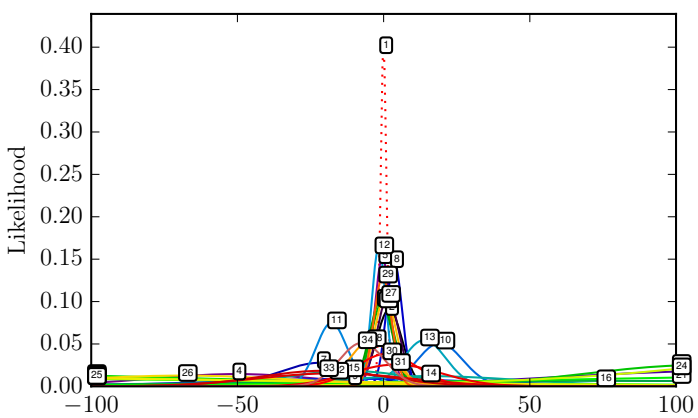
same δm_2
different m_2

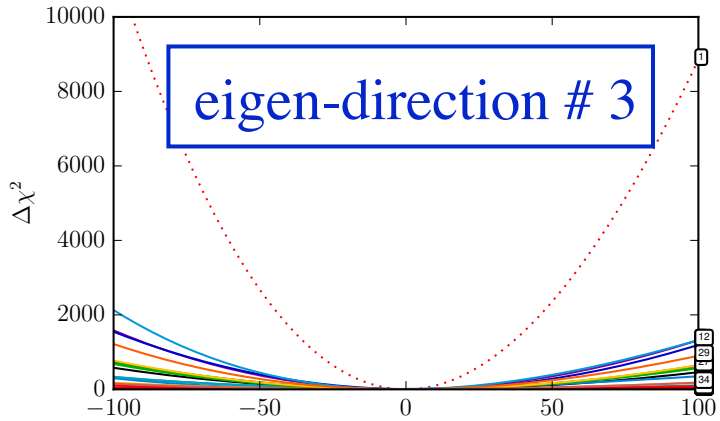
→ total uncertainty remains independent of degree of (in)compatibility of data

→ Gaussian likelihood gives unrealistic representation of true uncertainty

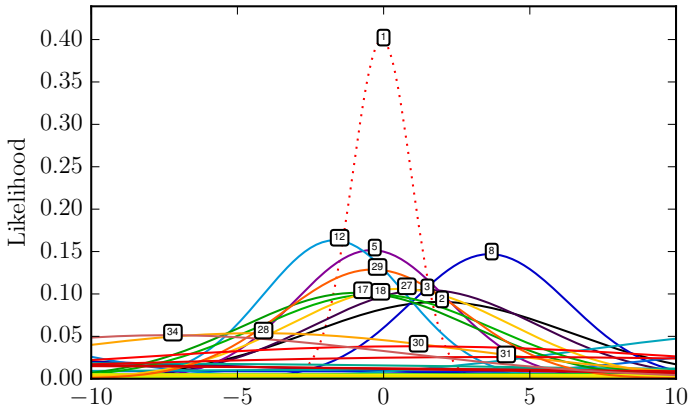
Incompatible data sets

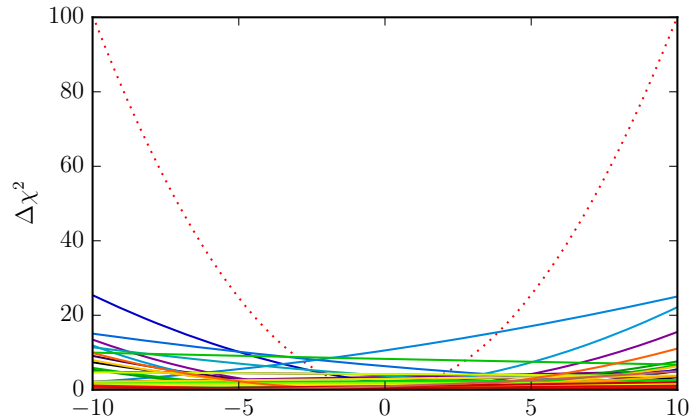
- Realistic example: recent CJ (CTEQ-JLab) global PDF analysis



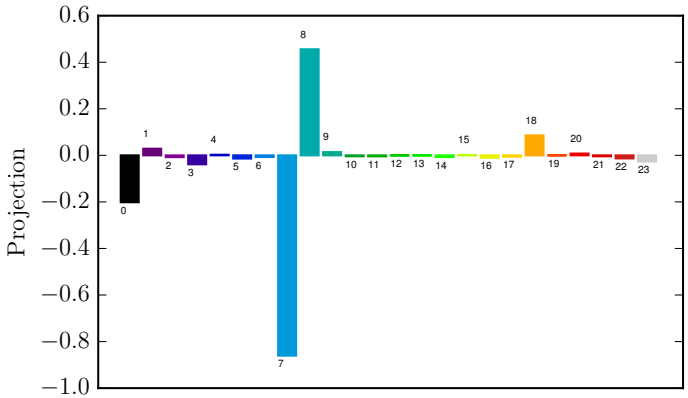


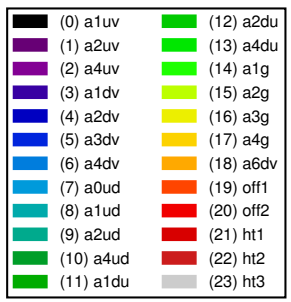
→ 24 parameters, 33 data sets

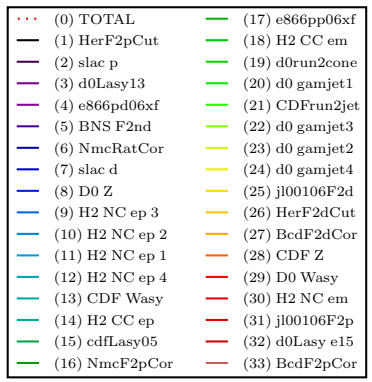




→ data sets compatible along this e-direction

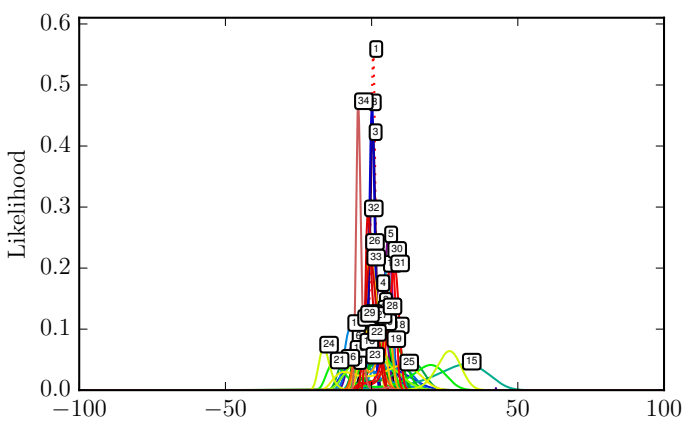


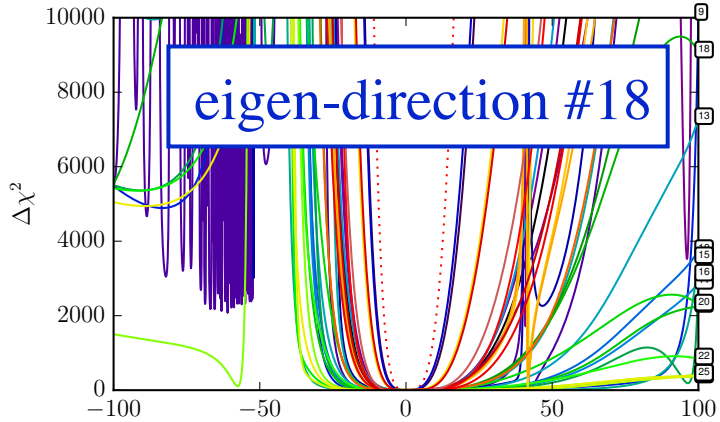




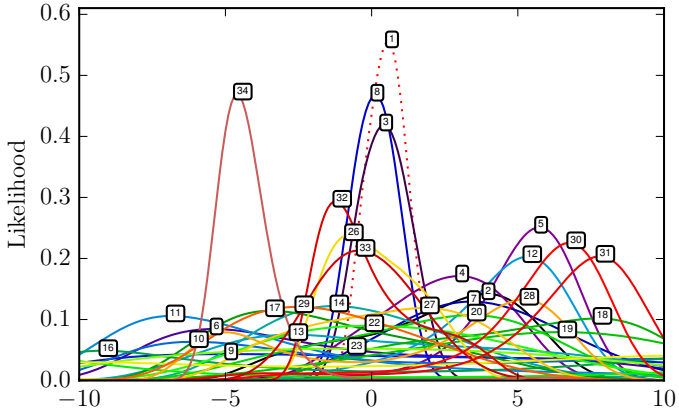
Incompatible data sets

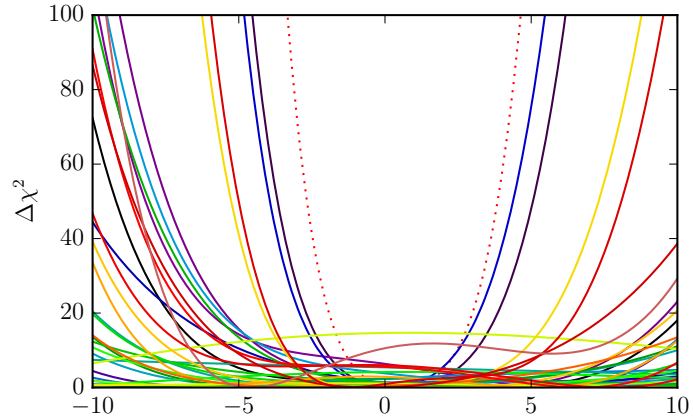
- Realistic example: recent CJ (CTEQ-JLab) global PDF analysis



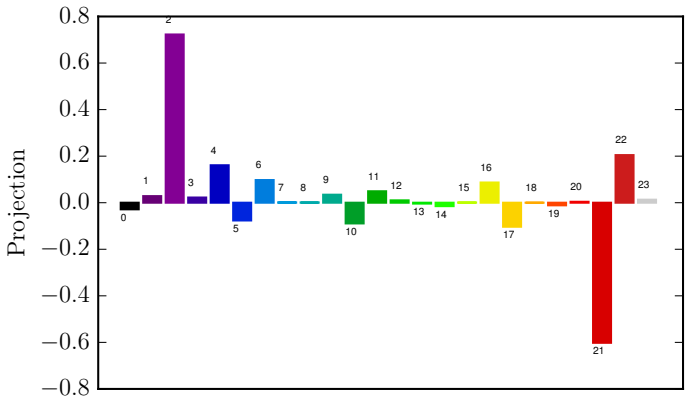


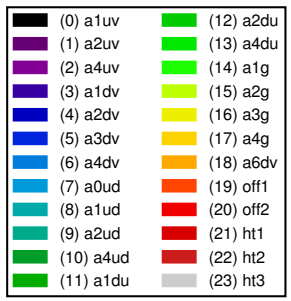
→ 24 parameters, 33 data sets

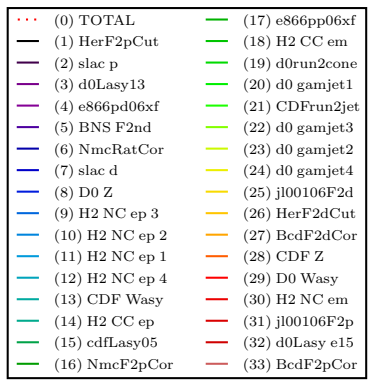




→ data sets not compatible along this e-direction

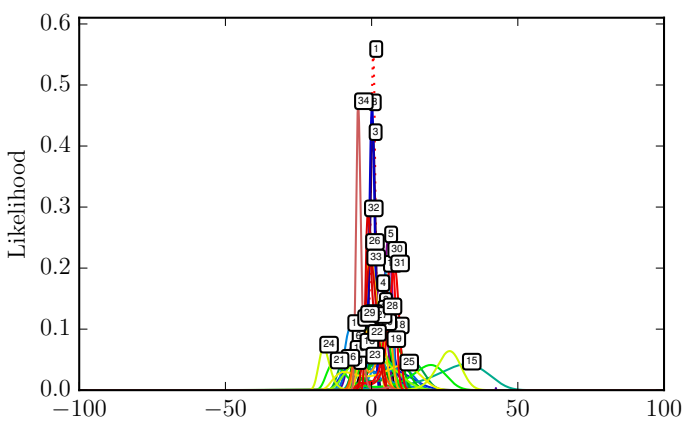


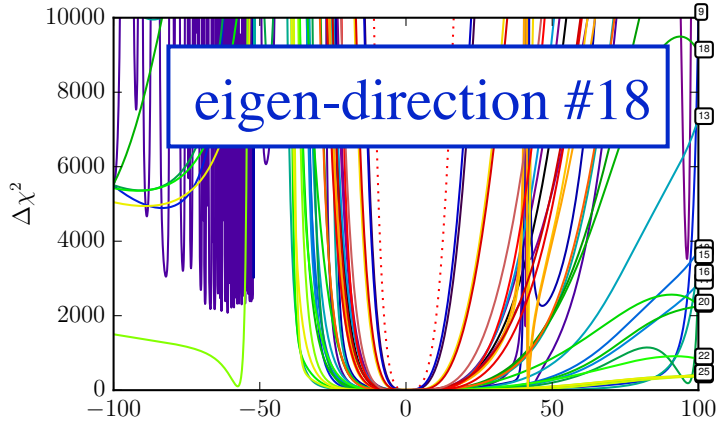




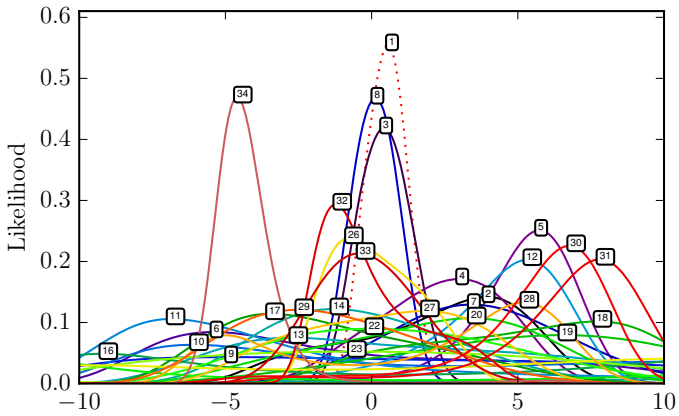
Incompatible data sets

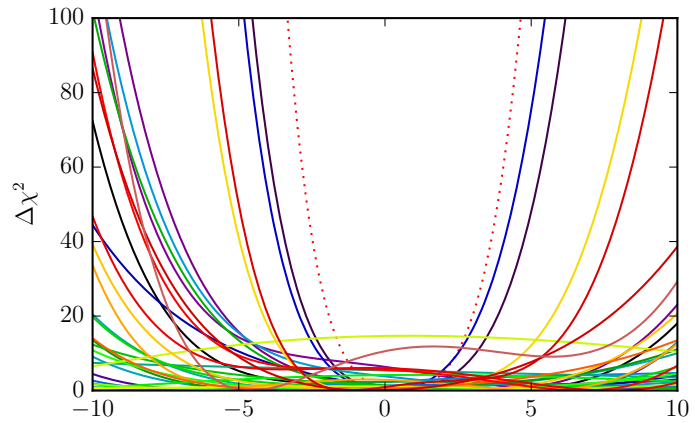
- Realistic example: recent CJ (CTEQ-JLab) global PDF analysis





→ 24 parameters, 33 data sets



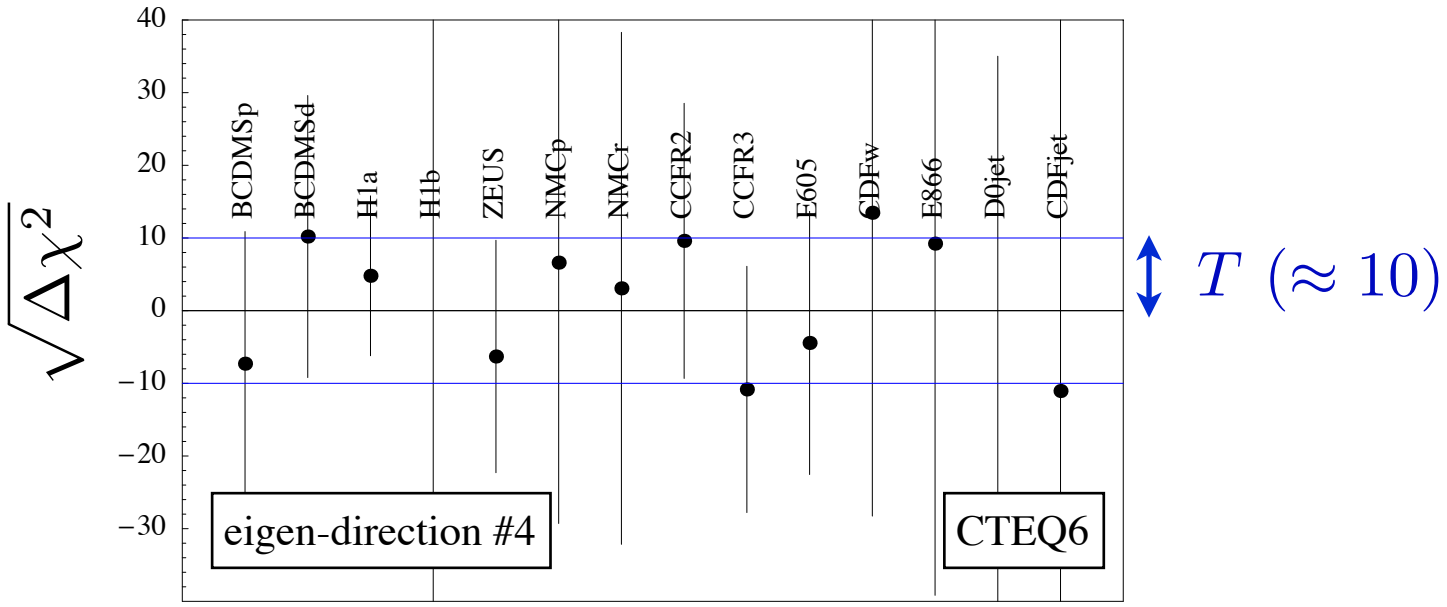


→ data sets not compatible along this e-direction

→ standard Gaussian likelihood incapable of accounting for underestimated individual errors (leading to incompatible data sets)
— not designed for such scenarios!

Incompatible data sets

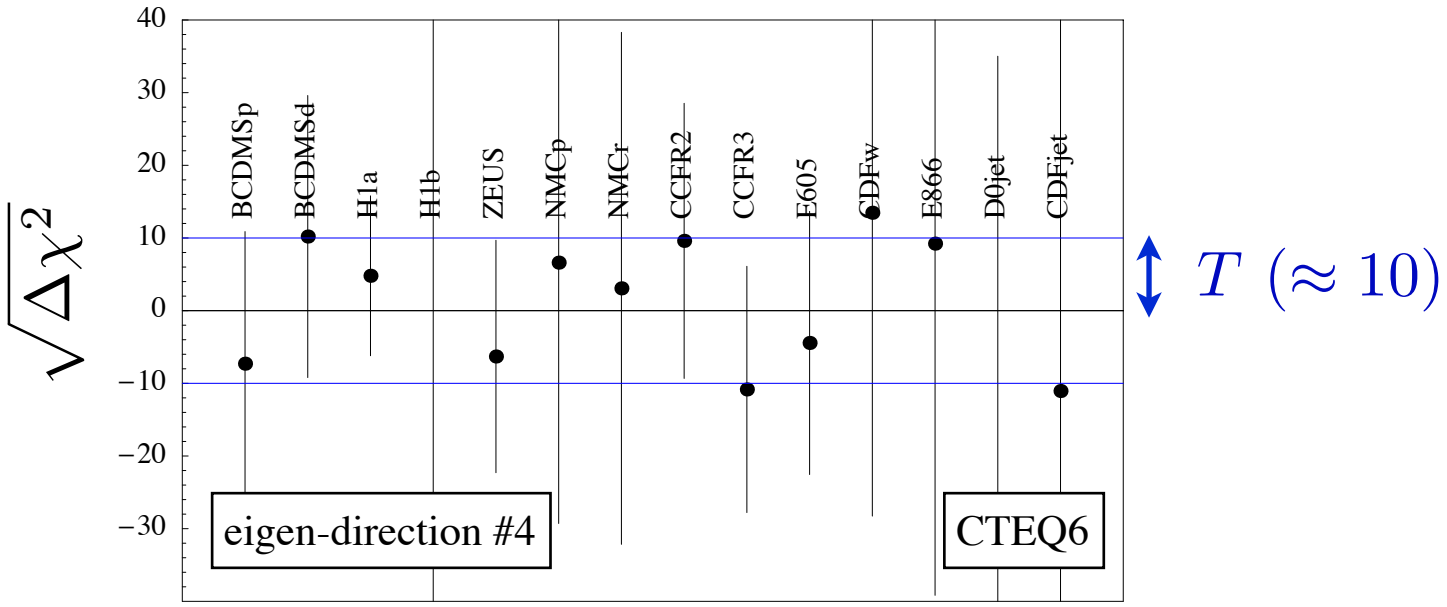
CTEQ tolerance criteria



- for each experiment, find minimum χ^2 along given e-direction
- from χ^2 distribution determine 90% CL for each experiment
- along each side of e-direction, determine maximum range $d_k^{\pm}$ allowed by the most constraining experiment
- T computed by averaging over all $d_k^{\pm}$ (typically $T \sim 5 - 10$)

Incompatible data sets

- CTEQ tolerance criteria



- This approach is *not consistent* with Gaussian likelihood

  → no clear Bayesian interpretation of uncertainties (ultimately, a prescription…)

To summarize standard maximum likelihood method…

- Gradient search (in parameter space) depends how "good" the starting point is

  ⟶ for ~30 parameters trying different starting points is impractical, if do not have some information about shape

- Common to free parameters initially, then freeze those not sensitive to data (χ^2 flat locally)

  ⟶ introduces bias, does not guarantee that flat χ^2 globally

- Cannot guarantee solution is unique

- Error propagation characterized by quadratic χ^2 near minimum

  ⟶ no guarantee this is quadratic globally (*e.g.* Student *t*-distribution?)

- Introduction of tolerance modifies Gaussian statistics

Monte Carlo

- Designed to faithfully compute Bayesian master formulas
- Do not assume a single minimum, include all possible solutions (with appropriate weightings)
- Do not assume likelihood is Gaussian in parameters
- Allows likelihood analysis to be extended to address tensions among data sets via Bayesian inference
- More computationally demanding compared with Hessian method

Monte Carlo

- First group to use MC for global PDF analysis was NNPDF, using neural network to parametrize $P(x)$ in

  Forte et al. (2002)

$$f(x) = N\, x^{\alpha}(1-x)^{\beta}\, P(x)$$

  — α, β are fitted "preprocessing coefficients"

- Iterative Monte Carlo (IMC), developed by JAM Collaboration, variant of NNPDF, tailored to non-neutral net parametrizations

- Markov Chain MC (MCMC) / Hybid MC (HMC)
  — recent "proof of principle" analysis, ideas from lattice QCD

  Gbedo, Mangin-Brinet (2017)

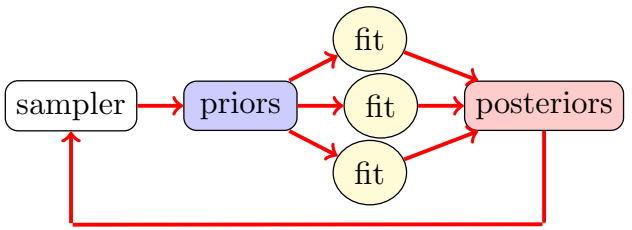
- Nested sampling (NS) — computes integrals in Bayesian master formulas (for $E,\ V,\ Z$) explicitly

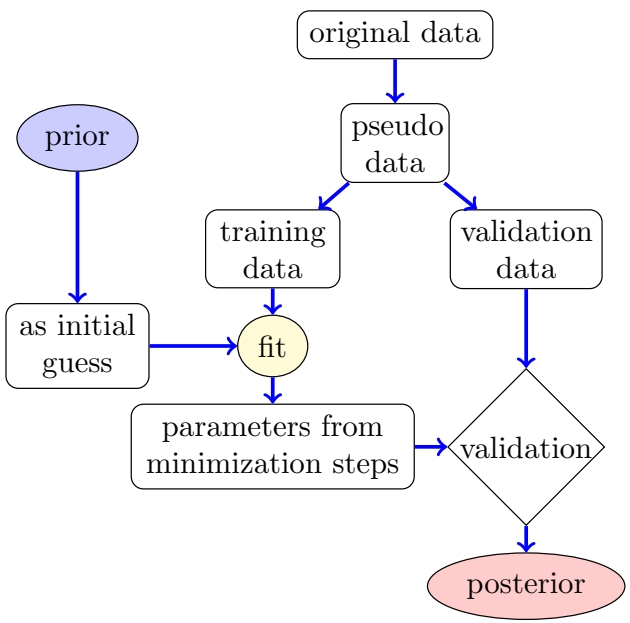
  Skilling (2004)

Iterative Monte Carlo (IMC)

- Use traditional functional form for input distribution shape, but sample significantly larger parameter space than possible in single-fit analyses

Iterative Monte Carlo (IMC)





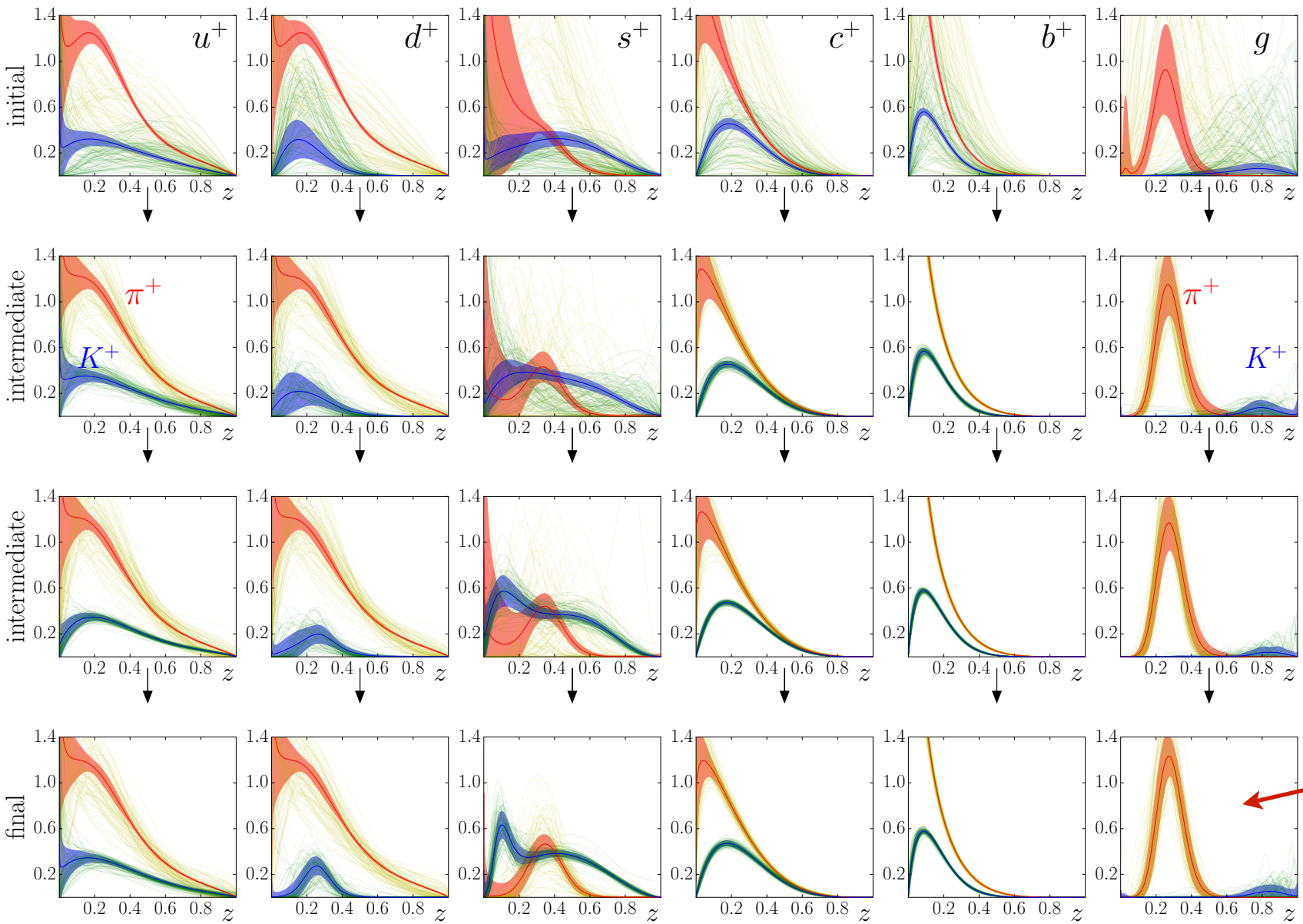
→ no assumptions for exponents

→ cross-validation to avoid overfitting

→ iterate until convergence criteria satisfied

Iterative Monte Carlo (IMC)

- *e.g.* of convergence (for fragmentation functions) in IMC



~ 20 iterations

Sato, Ethier, WM, Hirai, Kumano, Accardi (2016)

Incompatible data sets

- Rigorous (Bayesian) way to address incompatible data sets is to use generalization of Gaussian likelihood
  - joint vs. disjoint distributions
  - empirical Bayes
  - hierarchical Bayes
  - others, used in different fields

PDFs in lattice QCD

- Recent progress in extracting x dependence of PDFs in lattice QCD from matrix element of nonlocal operator

$$\mathcal{M}_q(z,P_z) = \langle N(P_z)|\overline{\psi}_q(0,z)\gamma_z \mathcal{W}(z,0)\psi_q(0,0)|N(P_z)\rangle$$

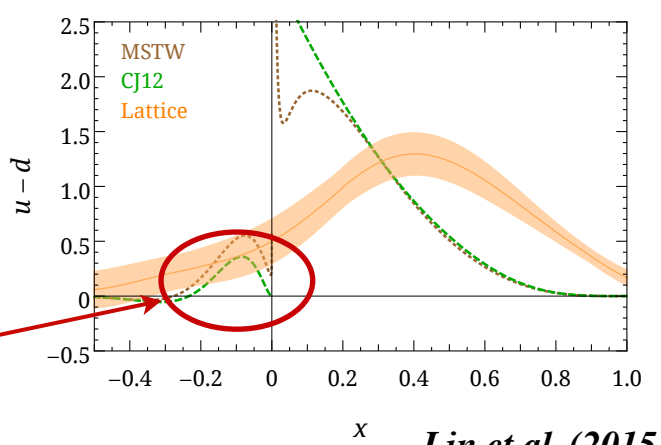
$$= \int_{-\infty}^{\infty} dy\ e^{iyP_z z}\ \widetilde{q}(y,P_z)$$

$\longrightarrow$ quasi-PDF $\widetilde{q}$ related to light-cone PDF via matching kernel $\widetilde{C}$

$$q(x,\mu) = \int_{-\infty}^{\infty} \frac{dy}{|y|}\ \widetilde{C}\left(\frac{x}{y},\mu,P_z\right)\ \widetilde{q}(y,P_z,\mu)$$

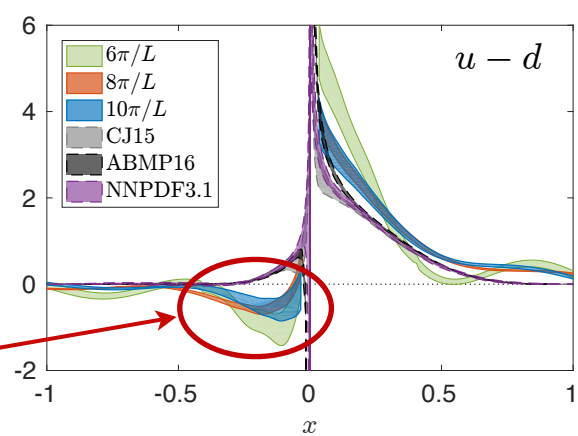
- Conflicting results on sign of $\bar{d}-\bar{u}$ asymmetry

$\bar{d} > \bar{u}$



Lin et al. (2015)

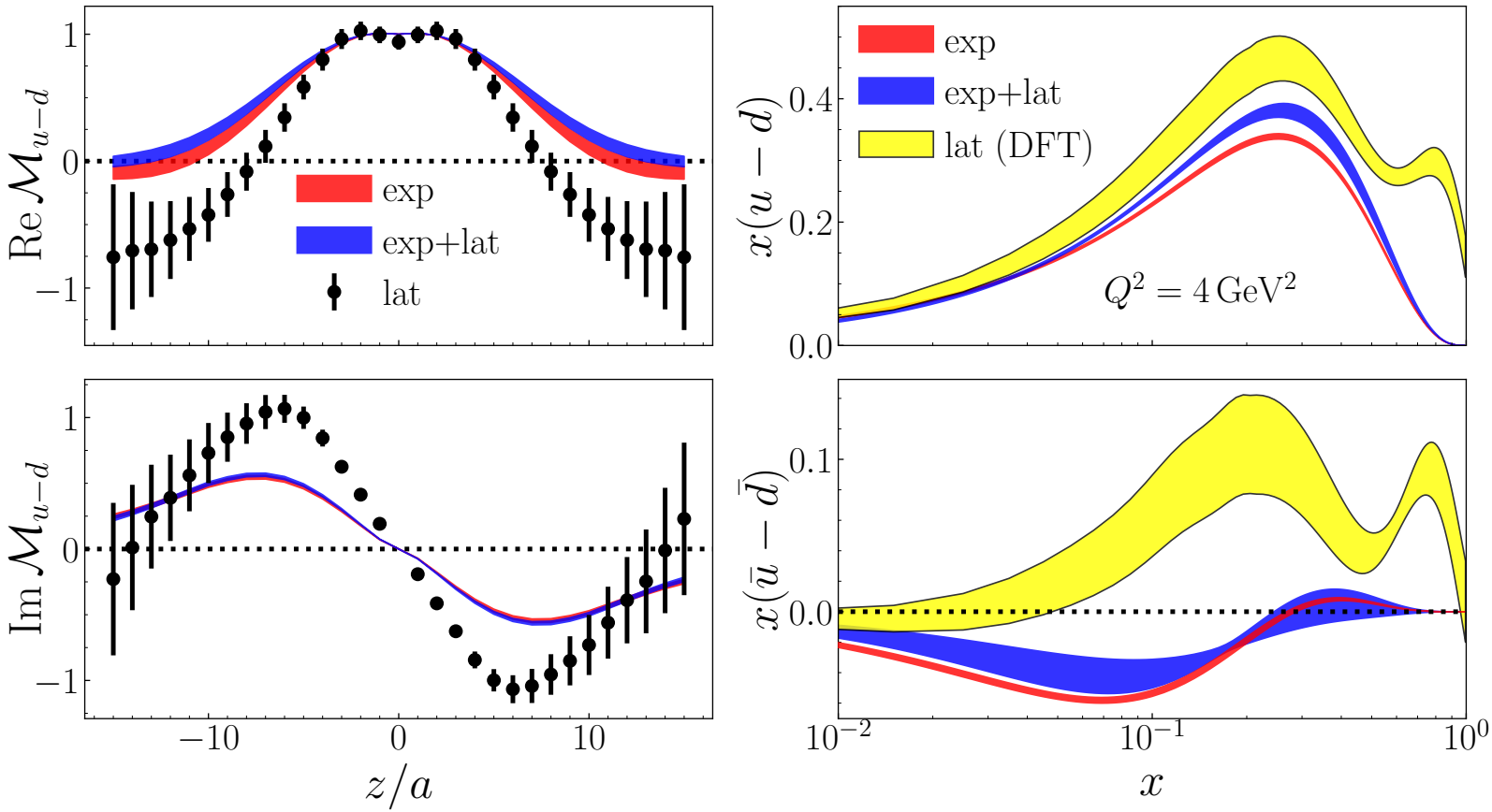
$\bar{d} < \bar{u}$



Alexandrou et al. (2018)

PDFs in lattice QCD

- Fit lattice observable directly within JAM framework



Bringewatt, Sato, WM, Qiu, Steffens, Constantinou (2021)

→ reatively weak impact of present lattice data on unpolarized PDF determination

PDFs in lattice QCD

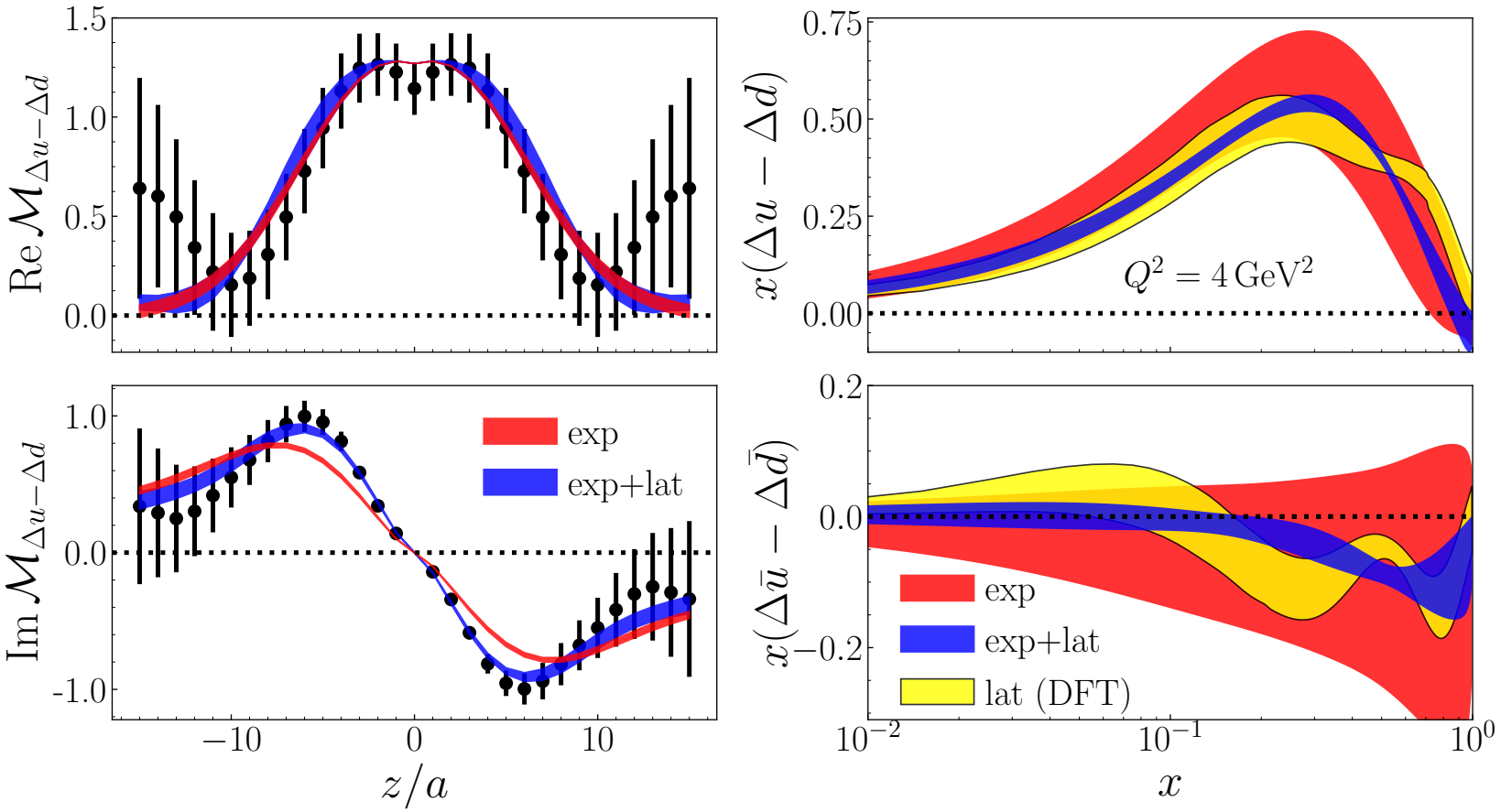
- Fit lattice observable directly within JAM framework

Observable	# data pts	χ^2/datum exp	χ^2/datum exp +lat
BCDMS F_2^p [23]	348	1.1	1.1
BCDMS F_2^d [24]	254	1.1	1.2
SLAC F_2^p [25]	218	1.4	1.4
SLAC F_2^d [25]	228	1.0	1.1
NMC F_2^p [26]	273	1.9	1.9
NMC F_2^d/F_2^p [27]	174	1.1	1.2
HERA $\sigma_{\rm NC}^{e^+p}$ (1) [30]	402	1.6	1.6
HERA $\sigma_{\rm NC}^{e^+p}$ (2) [30]	75	1.2	1.2
HERA $\sigma_{\rm NC}^{e^+p}$ (3) [30]	259	1.0	1.0
HERA $\sigma_{\rm NC}^{e^+p}$ (4) [30]	209	1.1	1.1
HERA $\sigma_{\rm NC}^{e^-p}$ [30]	159	1.7	1.7
HERA $\sigma_{\rm CC}^{e^+p}$ [30]	39	1.4	1.2
HERA $\sigma_{\rm CC}^{e^-p}$ [30]	42	1.4	1.4
E866 $\sigma_{\rm DY}^{pp}$ [28]	121	1.3	1.3
E866 $\sigma_{\rm DY}^{pd}$ [28]	129	1.7	1.8
ETMC19 Re$\mathcal{M}_{u-d}$ [8]	31		4.7
ETMC19 Im$\mathcal{M}_{u-d}$ [8]	30		22.7
Total (exp)	2,930	1.3	
(exp +lat)	2,991		1.6

→ difficult to fit current lattice matrix element data for unpolarized $u-d$ PDFs

PDFs in lattice QCD

- Fit lattice observable directly within JAM framework



Bringewatt, Sato, WM, Qiu, Steffens, Constantinou (2021)

→ better agreement between lattice and experiment for polarized PDFs (within larger uncertainties)

PDFs in lattice QCD

- Fit lattice observable directly within JAM framework

Observable	# data pts	χ^2/datum exp	χ^2/datum exp +lat
EMC A_1^p [31]	10	0.3	0.3
SMC A_1^p [32]	11	0.6	0.7
SMC A_1^d [32]	11	2.4	2.3
SMC A_1^p [33]	7	1.3	1.3
SMC A_1^d [33]	7	0.7	0.7
COMPASS A_1^p [34]	11	1.0	0.9
COMPASS A_1^d [35]	11	0.5	0.5
COMPASS A_1^p [36]	35	1.0	1.0
SLAC E80/E130 $A_\parallel^p$ [37]	10	0.8	0.8
SLAC E143 $A_\parallel^p$ [39]	39	0.9	0.8
SLAC E143 $A_\parallel^d$ [39]	39	1.0	1.0
SLAC E143 $A_\perp^p$ [39]	33	1.0	1.0
SLAC E143 $A_\perp^d$ [39]	33	1.2	1.2
SLAC E155 $A_\parallel^p$ [41]	59	1.5	1.4
SLAC E155 $A_\parallel^p$ [42]	59	1.1	1.1
SLAC E155 $A_\perp^p$ [43]	46	0.8	0.8
SLAC E155 $A_\perp^d$ [43]	46	1.5	1.5
SLAC E155x $\tilde{A}_\perp^p$ [44]	69	1.3	1.3
SLAC E155x $\tilde{A}_\perp^d$ [44]	69	0.9	0.9
HERMES A_1^n [45]	5	0.3	0.3
HERMES A_1^p [46]	16	0.6	0.6
HERMES A_1^p [46]	16	1.3	1.3
HERMES A_2^p [47]	9	1.1	1.1
ETMC19 Re$\mathcal{M}_{\Delta u-\Delta d}$ [8]	31		0.5
ETMC19 Im$\mathcal{M}_{\Delta u-\Delta d}$ [8]	30		0.3
Total (exp)	651	1.1	
(exp +lat)	712		1.0

→ good fits possible to both experimental and lattice data for polarized $u-d$ PDFs

Outlook

- New approaches being developed for global QCD analysis
  - simultaneous determination of parton distributions using Monte Carlo sampling of parameter space
  - towards synthesis of lattice QCD data with global analysis

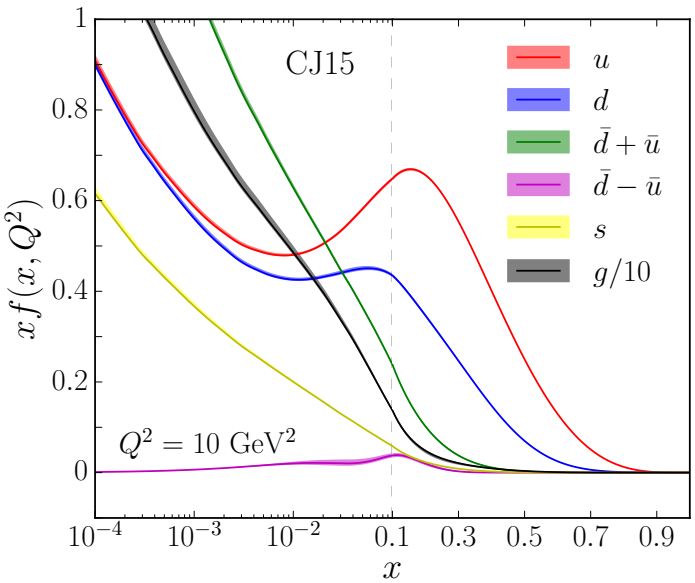
- Treatment of discrepant data sets needs attention
  - Bayesian perspective has clear merits

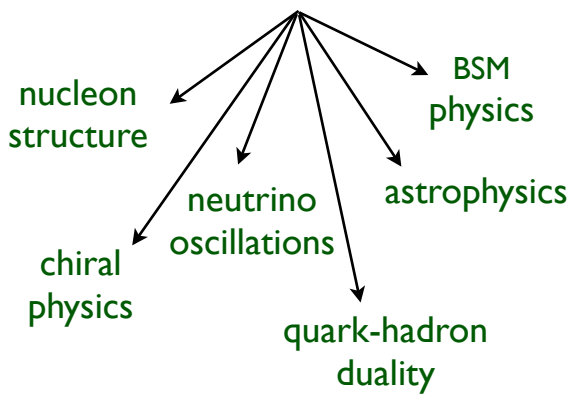
- Near-term future: "universal" QCD analysis of all observables sensitive to collinear (unpolarized & polarized) PDFs and FFs

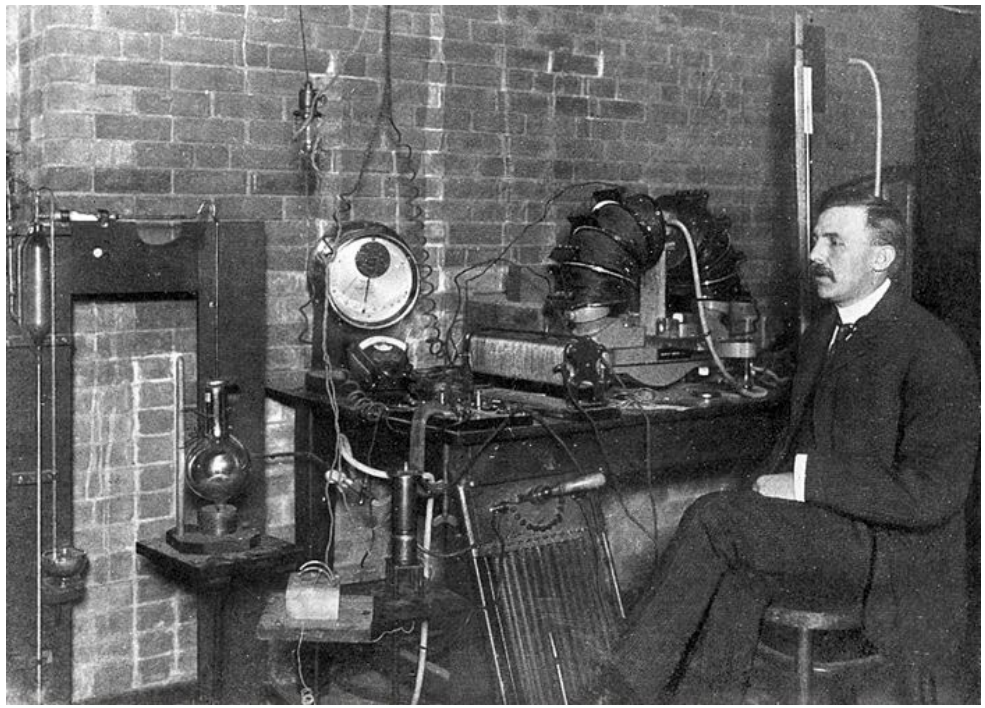
- Longer-term: apply MC technology to global QCD analysis of transverse momentum dependent (TMD) PDFs and FFs

Outlook

- Study of PDFs has brought together essential elements of nuclear and high-energy physics







Disjoint distributions

- Instead of using total likelihood that is a product ("and") of individual likelihoods, *e.g.* for simple example of two measurements

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

use instead sum ("or") of individual likelihoods

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

→ gives rather different expectation value and variance

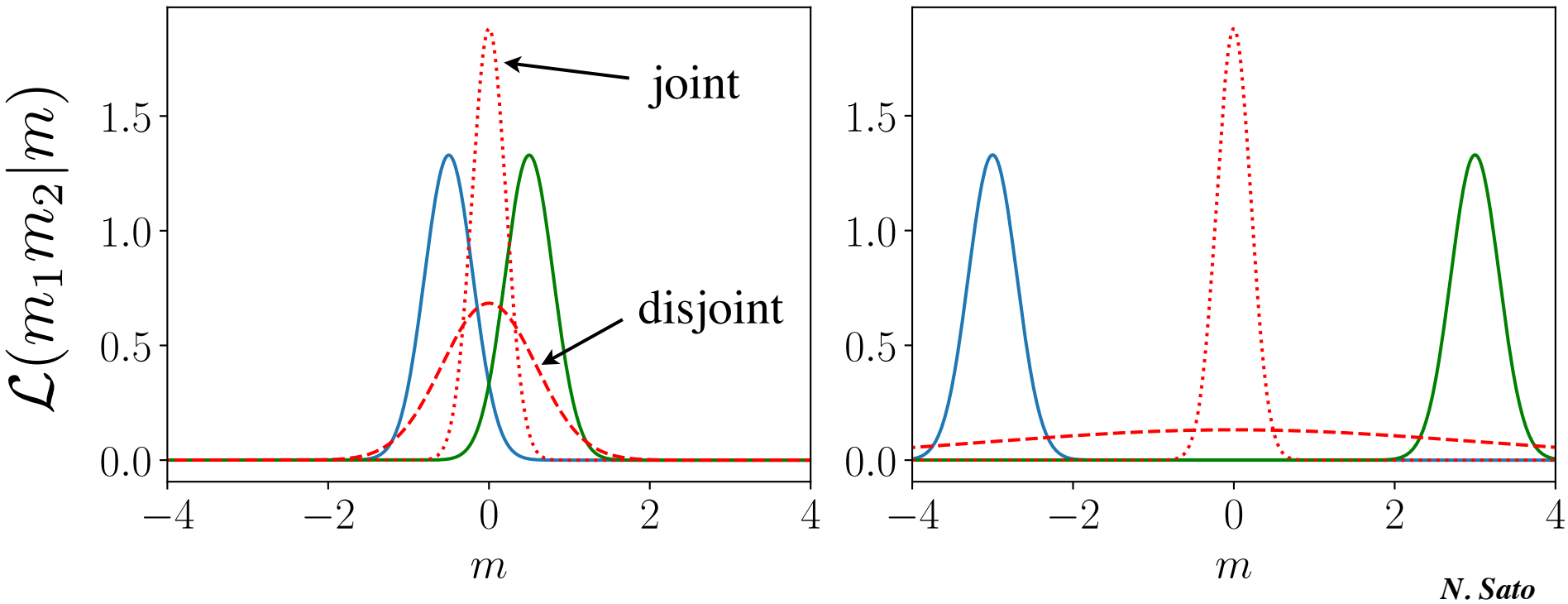
$$E[m] = \frac{1}{2}(m_1 + m_2)$$

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

depends on separation!

Disjoint distributions

- Symmetric uncertainties $\delta m_1 = \delta m_2$

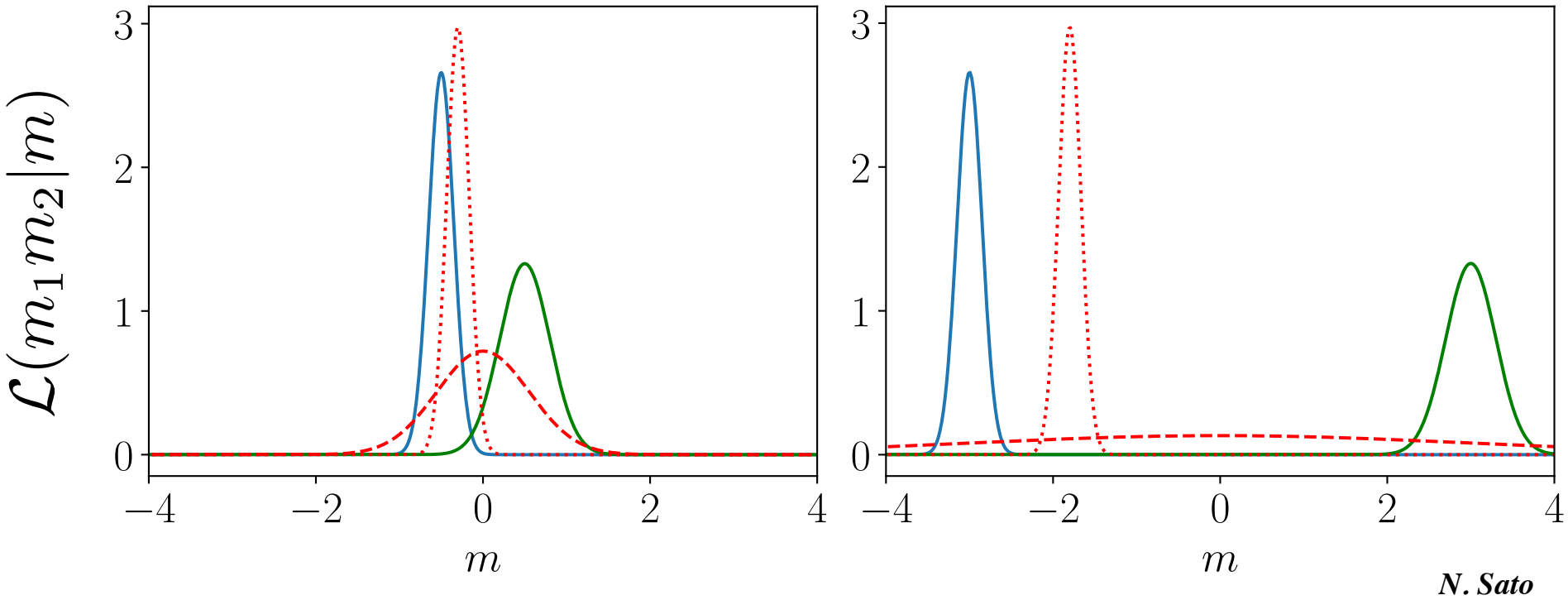


disjoint: $V[m] = \frac{1}{2}(\delta m_1^2 + \delta m_2^2) + \left(\frac{m_1 - m_2}{2}\right)^2$

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

Disjoint distributions

- Asymmetric uncertainties $\delta m_1 \neq \delta m_2$



→ disjoint likelihood gives broader overall uncertainty, overlapping individual (discrepant) data

Empirical Bayes

- Shortcoming of conventional Bayesian — still assume prior distribution follows specific form (*e.g.* Gaussian)

- Extend approach to more fully represent prior uncertainties, with final uncertainties that do not depend on initial choices

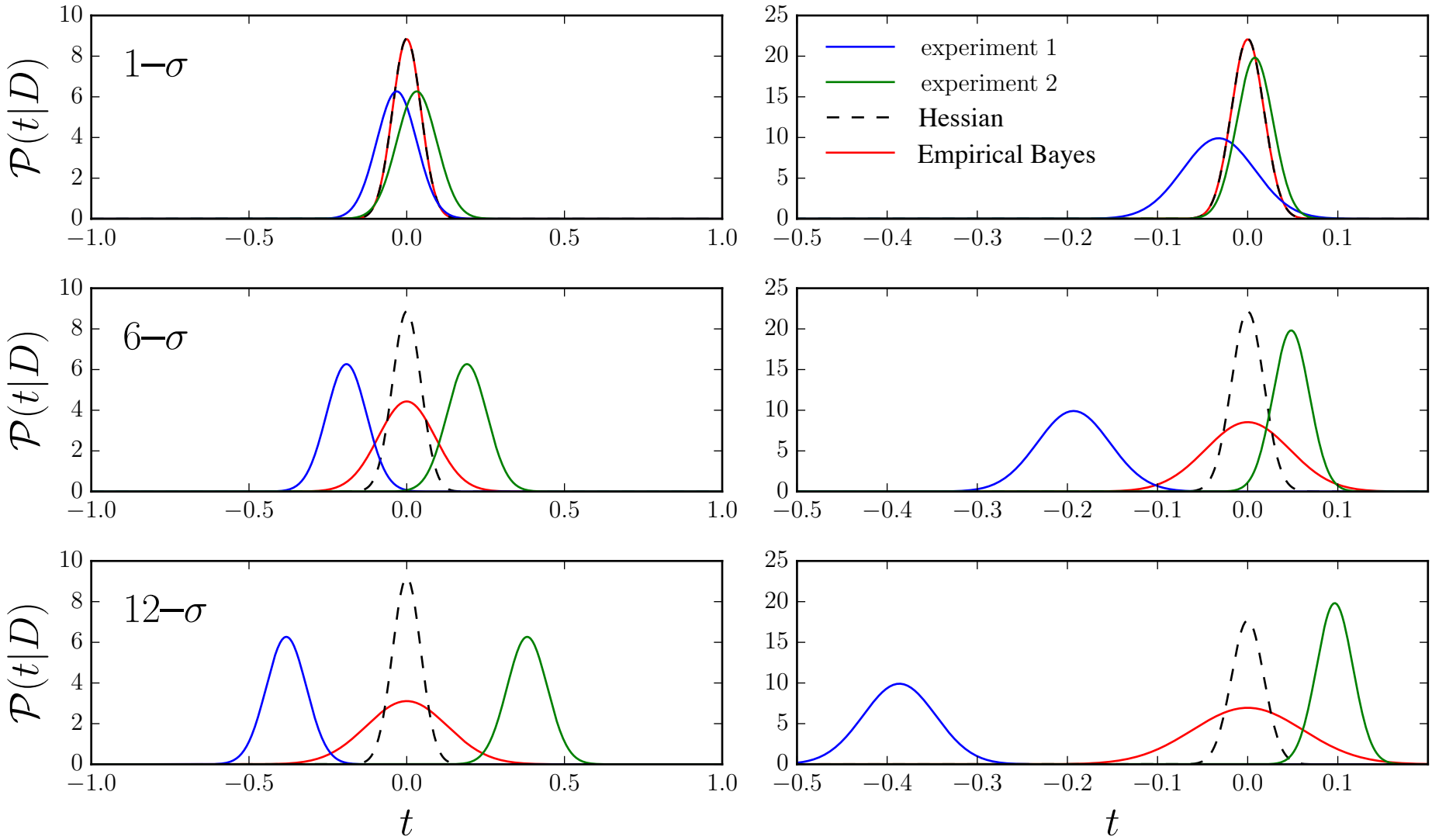
- In generalized approach, data uncertainties modified by distortion parameters, whose probability distributions given in terms of "hyperparameters" (or "nuisance parameters")

- Hyperparameters determined from data

  → give posteriors for both PDF and hyperparameters

Empirical Bayes

- Simple example of EB for symmetric & asymmetric errors



N. Sato