PDF Errors with Normalization Uncertainties

- Hessian vs Monte Carlo
- The D'Agostini Bias
- Bias free Fitting

NNPDF: RDB, Luigi Del Debbio, Stefano Forte,
Alberto Guffanti, Jose Latorre, Juan Rojo, Maria Ubiali
(Barcelona, Edinburgh, Freiburg, Milan)

PDF4LHC
CERN August 2009

(NN)PDFs for LHC

To fully exploit LHC data, we need:

- Precise reliable faithful PDFs
- No theoretical bias (beyond NLO pQCD, etc.)

  No bias due to functional form

  No bias due to improper statistical procedure

- Genuine statistical confidence level

  Full inclusion of correlations in exp systematics

  Full inclusion of normalization uncertainties

  No rescaling of experimental errors

  Uniform treatment of uncertainties

(NN)PDFs for LHC

To fully exploit LHC data, we need:

- Precise reliable faithful PDFs
- No theoretical bias (beyond NLO pQCD, etc.)
  - No bias due to functional form
  - No bias due to improper statistical procedure
- Genuine statistical confidence level
  - Full inclusion of correlations in exp systematics
  - Full inclusion of normalization uncertainties
  - No rescaling of experimental errors
  - Uniform treatment of uncertainties

Zero Tolerance!

Catalogue of (average) Errors

	Process	Expt Set	N_{data}	Stat	Syst	**Norm**	
	Fixed Target	SLACp	211	2.7%		2.2%	
		BCDMSp	351	3.2%	2.0%	3.2%	
		NMCp	288	3.7%	2.3%	2.0%	S
	HERA	Z97NC	160	6.2%	3.1%	2.1%	M
		Z02NC	92	12.7%	2.3%	1.8%	A
1.0		Z03NC	90	7.7%	3.3%	2.0%	L
		Z06NC	90	3.8%	3.7%	2.6%	L
		H197NC	130	12.5%	3.2%	1.5%	
		H199NC	126	14.9%	2.8%	1.8%	
		H100NC	147	9.4%	3.2%	1.5%	
	Nu-DIS	CHORUSnu	607	4.2%	6.4%	2.1%	
1.2		NTVnuDMN	45	17.2%	1.0%	2.2%	
	Drell-Yan	DYE605	119	12.7%	9.9%	14.9%	
		DYE886p	184	19.9%	5.0%	6.5%	B
2.0	Incl Jets	CDFR2KT	76	4.5%	21.1%	5.7%	I
		D0R2CON	110	3.6%	14.9%	6.1%	G

Hessian Method

cov_0

Simple model: one observable t, two data points $m_1 \pm \sigma_1$ $m_2 \pm \sigma_2$:

$$\chi^2 = \frac{t - m_1)^2}{\sigma_1^2} + \frac{(t - m_2)^2}{\sigma_2^2}$$

Minimise:

$$\boxed{t = \left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right)\Sigma^2 \equiv w}$$

Variance:

$$V_{tt} = \left(\frac{1}{2}\frac{\partial^2 \chi^2}{\partial t^2}\right)^{-1} = \Sigma^2 \qquad \boxed{\frac{1}{\Sigma^2} \equiv \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$$

Unbiased: if $\sigma_1 = \sigma_2 = \sigma$: $\quad t = \frac{1}{2}(m_1 + m_2) \equiv \bar{m}, \quad V_{tt} = \frac{\sigma^2}{2}$

Decoupling: if $\sigma_2 \gg \sigma_1$: $\quad t \sim m_1, \quad V_{tt} \sim \sigma_1^2$

Monte Carlo Method

Generate data "replicas": Gaussian random variables M_1, M_2

$$\langle M_i \rangle = m_i \qquad \langle M_i^2 \rangle = m_i^2 + \sigma_i^2$$

Fit to each replica:

$$\chi^2(T) = \frac{(T - M_1)^2}{\sigma_1^2} + \frac{(T - M_2)^2}{\sigma_2^2}$$

Minimise:

$$T = \left(\frac{M_1}{\sigma_1^2} + \frac{M_2}{\sigma_2^2} \right) \Sigma^2$$

Then

$$\begin{aligned} \mathrm{E}[t] &\equiv \langle T \rangle = w \\ \mathrm{Var}[t] &\equiv \langle T^2 \rangle - (\langle T \rangle)^2 = \Sigma^2 \end{aligned}$$

Same results as Hessian: unbiased etc.

Hessian $\equiv$ Monte Carlo

(for additive Gaussian uncertainties)

Normalization: 1 expt

cov_0

Two data points $m_1 \pm \sigma_1$ and $m_2 \pm \sigma_2$ from single expt.
Overall multiplicative normalization uncertainty $1 \pm s$.
Replicas: (NM_1, NM_2) with N another Gaussian random variable:

$$\langle N \rangle = 1 \qquad \langle N^2 \rangle = 1 + s^2 \qquad \langle N^n M_i^n \rangle = \langle N^n \rangle \langle M_i^n \rangle$$

Fit to replica depends only on ratio σ_1/σ_2, not on s: take

$$\chi^2(T) = \frac{(T - NM_1)^2}{\sigma_1^2} + \frac{(T - NM_2)^2}{\sigma_2^2}$$

$$T = N \left(\frac{M_1}{\sigma_1^2} + \frac{M_2}{\sigma_2^2} \right) \Sigma^2$$

$$\begin{aligned} \text{E}[t] &\equiv \langle T \rangle = w \\ \text{Var}[t] &\equiv \langle T^2 \rangle - (\langle T \rangle)^2 = \Sigma^2 (1 + s^2) + s^2 w^2 \end{aligned}$$

Factor $1 + s^2$:

$$\text{Var}[NT] = \text{E}[N]\text{Var}[T] + \text{E}[T]\text{Var}[N] + \text{Var}[N]\text{Var}[T]$$

Normalization: 2 expt

cov_0

Data points $m_1 \pm \sigma_1$ and $m_2 \pm \sigma_2$ from two independent expt. with multiplicative normalization uncertainties $1 \pm s_1$ $1 \pm s_2$.
Replicas: $(N_1 M_1, N_2 M_2)$ with N_1 N_2 Gaussian random variables:

$$\langle N_i \rangle = 1 \qquad \langle N_i^2 \rangle = 1 + s_i^2 \qquad \langle N_1^{n_1} N_2^{n_2} \rangle = \langle N_1^{n_1} \rangle \langle N_2^{n_2} \rangle$$

Consider (NNPDF1.x)

$$\chi^2(T) = \frac{(T - N_1 M_1)^2}{\sigma_1^2} + \frac{(T - N_2 M_2)^2}{\sigma_2^2}$$

$$T = \left(\frac{N_1 M_1}{\sigma_1^2} + \frac{N_2 M_2}{\sigma_2^2} \right) \Sigma^2$$

$$\begin{aligned} \mathrm{E}[t] &\equiv \langle T \rangle = w \\ \mathrm{Var}[t] &\equiv \langle T^2 \rangle - (\langle T \rangle)^2 = \Sigma^2 (1 + s^2) + \Sigma^4 \sum_i s_i^2 \frac{m_i^2 + \sigma_i^2}{\sigma_i^4} \end{aligned}$$

Problem: if $s_2 \gg s_1$, E[t] unchanged: expt 2 does not decouple.
Need to include s_i in χ^2 weighting.

Including Norm. Errors in χ^2

(Problems)

Norm. in covariance: 1 expt

m-cov

Build χ^2 using covariance matrix:

$$(\mathrm{cov})_{ij} = \langle N^2 M_i M_j \rangle - \langle N M_i \rangle \langle N M_j \rangle = \sigma_i^2 \delta_{ij} + s^2 m_i m_j$$

$$\chi^2(t) = \sum_{ij} (t - m_i)(\mathrm{cov}^{-1})_{ij} (t - m_j)$$

$$t = \frac{w}{1 + r^2 s^2 w^2 / \Sigma^2} \qquad r^2 \equiv \Sigma^2 \sum_i \frac{m_i^2 - w^2}{\sigma_i^2}$$

$$V_{tt} = \frac{\Sigma^2 + s^2 w^2}{1 + r^2 s^2 w^2 / \Sigma^2}$$

Problem: downward bias: if $\sigma_1 = \sigma_2$, $t = \bar{m}/(1 + 2r^2 s^2 \bar{m}^2/\sigma^2)$
For N data

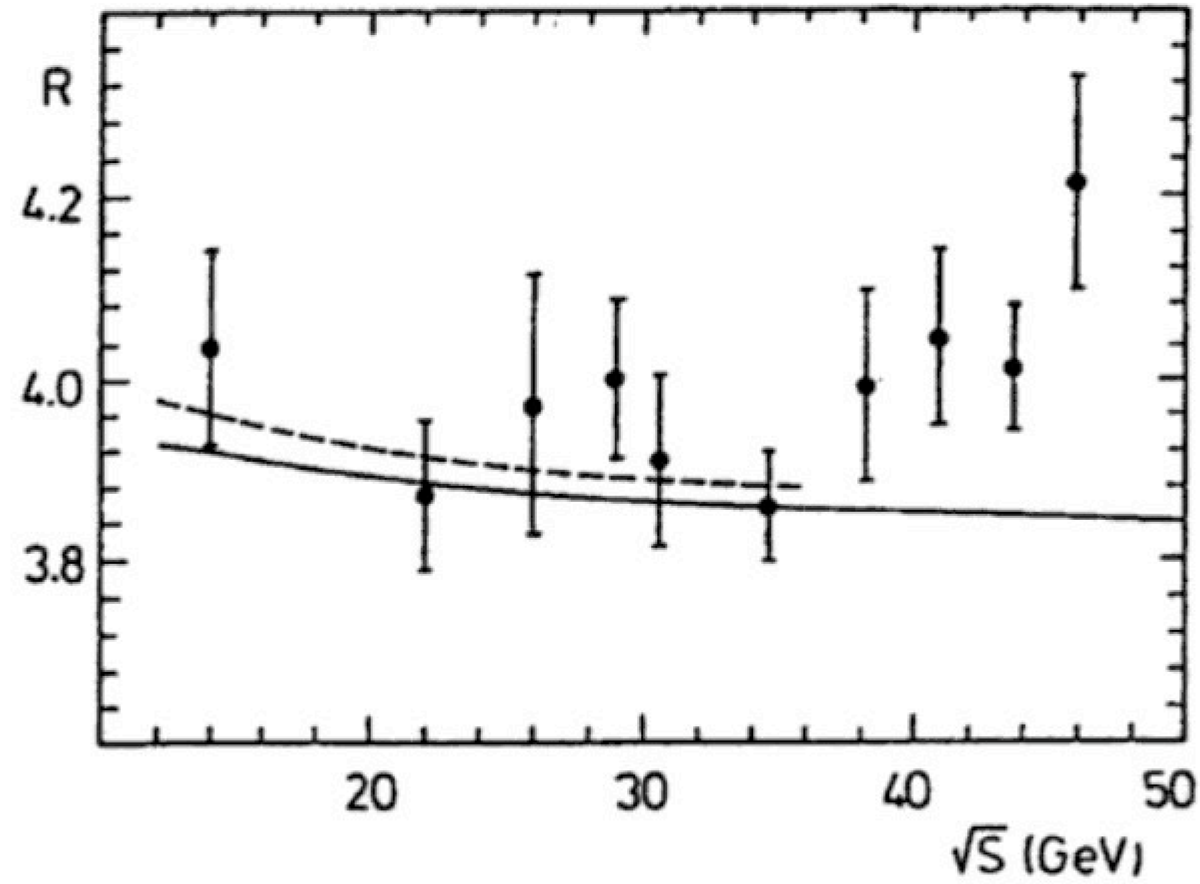
$$t = \bar{m}/(1 + \mathrm{N} r^2 s^2 \bar{m}^2/\sigma^2) \sim \bar{m}/(1 + \mathrm{N} s^2)$$

DISASTER!

D'Agostini 1994

$R(e^+e^-)$

CELLO 1987



Dashed line: data below 36 GeV

Solid line: all data

D'Agostini 1994

Norm. in covariance: 2 expt

m-cov

$$(\text{cov})_{ij} = \langle N_i M_i N_j M_j \rangle - \langle N_i M_i \rangle \langle N_j M_j \rangle = (\sigma_i^2 + s_i^2 m_i^2)\delta_{ij}$$

$$\chi^2(t) = \frac{(t-m_1)^2}{\sigma_1^2 + s_1^2 m_1^2} + \frac{(t-m_2)^2}{\sigma_2^2 + s_2^2 m_2^2}$$

$$t = \sum_i \frac{m_i}{\sigma_i^2 + s_i^2 m_i^2} \Big/ \sum_i \frac{1}{\sigma_i^2 + s_i^2 m_i^2}$$

Decoupling: if $\sigma_1^2, \sigma_2^2 \ll s_1^2 m_1^2, s_2^2 m_2^2$

$$t = m_1 m_2 \frac{m_1/s_2^2 + m_2/s_1^2}{m_1^2/s_2^2 + m_2^2/s_1^2} \qquad V_{tt} = \frac{m_1^2 m_2^2}{m_1^2/s_2^2 + m_2^2/s_1^2}$$

So when $s_2 \gg s_1$, $t \sim m_1$, $V_{tt} \sim s_1^2 m_1^2$, but in peculiar way.

Norm. in covariance: 2 expt

m-cov

$$(\text{cov})_{ij} = \langle N_i M_i N_j M_j \rangle - \langle N_i M_i \rangle \langle N_j M_j \rangle = (\sigma_i^2 + s_i^2 m_i^2)\delta_{ij}$$

$$\chi^2(t) = \frac{(t - m_1)^2}{\sigma_1^2 + s_1^2 m_1^2} + \frac{(t - m_2)^2}{\sigma_2^2 + s_2^2 m_2^2}$$

$$t = \sum_i \frac{m_i}{\sigma_i^2 + s_i^2 m_i^2} \Big/ \sum_i \frac{1}{\sigma_i^2 + s_i^2 m_i^2}$$

Bias: (small) if $\sigma_1 = \sigma_2 = \sigma$, $s_1 = s_2 = s$, $r = \frac{m_1 - m_2}{m_1 + m_2}$

$$t = \bar{m}\left(1 - \tfrac{1}{2} r^2 \frac{s^2 \bar{m}^2}{\sigma^2 + s^2 \bar{m}^2} + \cdots\right)$$

"D'Agostini bias":
Fit prefers smaller data values because these have smaller errors.

D'Agostini 1994

The Penalty Trick: 1 expt

Treat normalization as fitted parameter n:

$$\chi^2(t,n) = \frac{(m_1 - t/n)^2}{\sigma_1^2} + \frac{(m_2 - t/n)^2}{\sigma_2^2} + \frac{(n-1)^2}{s^2}$$

Minimize w.r.t. t and n: gives

$$\begin{aligned} t &= nw \qquad\qquad n = 1 \\ V_{tt} &= \Sigma^2 + s^2 w^2 \end{aligned}$$

i.e. correct result, no bias.

N.B.

$$\chi^2(t,n) = \frac{(t - nm_1)^2}{\sigma_1^2} + \frac{(t - nm_2)^2}{\sigma_2^2} + \frac{(n-1)^2}{s^2}$$

gives same result as cov matrix: D'Agostini bias.

The Penalty Trick: 2 expt

Now have two normalizations to fit, n_1 and n_2:

$$\chi^2(t, n_i) = \frac{(m_1 - t/n_1)^2}{\sigma_1^2} + \frac{(m_2 - t/n_2)^2}{\sigma_2^2} + \frac{(n_1 - 1)^2}{s_1^2} + \frac{(n_2 - 1)^2}{s_2^2}$$

Minimize w.r.t. t, n_1 and n_2: gives

$$\begin{aligned} t &= \left(\frac{m_1}{n_1\sigma_1^2} + \frac{m_2}{n_2\sigma_2^2}\right) \Big/ \left(\frac{1}{n_1^2\sigma_1^2} + \frac{1}{n_2^2\sigma_2^2}\right) \\ n_i &= 1 + \frac{s_i^2 t}{n_i^2\sigma_i^2}\left(\frac{t}{n_i} - m_i\right) \end{aligned}$$

Three simultaneously **nonlinear** equations!

Decoupling: for $\sigma_1^2, \sigma_2^2 \ll s_1^2 m_1^2, s_2^2 m_2^2$:

$$t = m_1 m_2 \frac{m_1/s_2^2 + m_2/s_1^2}{m_1^2/s_2^2 + m_2^2/s_1^2} \qquad V_{tt} = m_1^2 m_2^2 \frac{1}{m_1^2/s_2^2 + m_2^2/s_1^2}$$

So decoupling works precisely as for m-cov.

The Penalty Trick: 2 expt

n-cov

Now have two normalizations to fit, n_1 and n_2:

$$\chi^2(t, n_i) = \frac{(m_1 - t/n_1)^2}{\sigma_1^2} + \frac{(m_2 - t/n_2)^2}{\sigma_2^2} + \frac{(n_1 - 1)^2}{s_1^2} + \frac{(n_2 - 1)^2}{s_2^2}$$

Minimize w.r.t. t, n_1 and n_2: gives

$$\begin{aligned} t &= \left(\frac{m_1}{n_1\sigma_1^2} + \frac{m_2}{n_2\sigma_2^2}\right) \Big/ \left(\frac{1}{n_1^2\sigma_1^2} + \frac{1}{n_2^2\sigma_2^2}\right) \\ n_i &= 1 + \frac{s_i^2 t}{n_i^2\sigma_i^2}\left(\frac{t}{n_i} - m_i\right) \end{aligned}$$

Three simultaneously **nonlinear** equations!

Bias: for $\sigma_1 = \sigma_2 = \sigma$, $s_1 = s_2 = s$, $r = \frac{m_1 - m_2}{m_1 + m_2}$:

$$t = \bar{m}\left(1 + \frac{s^2\bar{m}^2(\sigma^2 - 2s^2\bar{m}^2)}{(\sigma^2 + s^2\bar{m}^2)^2} r^2 + O(r^4)\right)$$

So biased when $m_1 \neq m_2$: D'Agostini bias (but nonlinear)
Bias small: eg $\sim 0.2\%$ at HERA; $\sim 1\%$ for DY and jets.

MSTW

Towards a Perfect χ^2

The Solution

Unbiased covariance : 1 expt t_0-cov

Use cov matrix, but take some t_0 instead of m_i:

$$(\text{cov})_{ij} = \sigma_i^2 \delta_{ij} + s^2 t_0^2$$

Then (in Monte Carlo method)

$$\chi^2(T) = \sum_{ij} (T - NM_i)(\text{cov}^{-1})_{ij}(T - NM_j)$$

$$T = N\Sigma^2 \left(\frac{M_1}{\sigma_1^2} + \frac{M_2}{\sigma_2^2} \right)$$

so

$$\begin{aligned} \text{E}[t] &= w \\ \text{Var}[t] &= \Sigma^2(1 + s^2) + s^2 w^2 \end{aligned}$$

Correct, unbiased, independent of t_0, as expected.

Unbiased covariance : 2 expt t_0-cov

$$(\mathrm{cov})_{ij} = (\sigma_i^2 + s_i^2 t_0^2)\delta_{ij}$$

$$\chi^2(T) = \sum_i \frac{(T - N_i M_i)^2}{\sigma_i^2 + s_i^2 t_0^2}$$

$$T = \Sigma_0^2 \sum_i \frac{N_i M_i}{\sigma_i^2 + s_i^2 t_0^2} \qquad \frac{1}{\Sigma_0^2} \equiv \sum_i \frac{1}{\sigma_i^2 + s_i^2 t_0^2}$$

$$\mathrm{E}[t] = \Sigma_0^2 \sum_i \frac{m_i}{\sigma_i^2 + s_i^2 t_0^2} \qquad \mathrm{Var}[t] = \Sigma_0^4 \sum_i \frac{\sigma_i^2 + s_i^2(m_i^2 + \sigma_i^2)}{(\sigma_i^2 + s_i^2 t_0^2)^2}$$

Unbiased: if $\sigma_1 = \sigma_2 = \sigma$, $s_1 = s_2 = s$

$$\mathrm{E}[t] = \bar{m} \qquad \mathrm{Var}[t] = \tfrac{1}{2}(\sigma^2(1 + s^2) + s^2 \tfrac{1}{2}(m_1^2 + m_2^2))$$

independent of t_0

Unbiased covariance : 2 expt

t_0-cov

$$(\text{cov})_{ij} = (\sigma_i^2 + s_i^2 t_0^2)\delta_{ij}$$

$$\chi^2(T) = \sum_i \frac{(T - N_i M_i)^2}{\sigma_i^2 + s_i^2 t_0^2}$$

$$T = \Sigma_0^2 \sum_i \frac{N_i M_i}{\sigma_i^2 + s_i^2 t_0^2} \qquad \frac{1}{\Sigma_0^2} \equiv \sum_i \frac{1}{\sigma_i^2 + s_i^2 t_0^2}$$

$$\text{E}[t] = \Sigma_0^2 \sum_i \frac{m_i}{\sigma_i^2 + s_i^2 t_0^2} \qquad \text{Var}[t] = \Sigma_0^4 \sum_i \frac{\sigma_i^2 + s_i^2(m_i^2 + \sigma_i^2)}{(\sigma_i^2 + s_i^2 t_0^2)^2}$$

Decoupling: if $\sigma_1^2, \sigma_2^2 \ll s_1^2 m_1^2, s_2^2 m_2^2$

$$\text{E}[t] = \frac{m_1/s_1^2 + m_2/s_2^2}{1/s_1^2 + 1/s_2^2} \qquad \text{Var}[t] = \frac{m_1^2/s_1^2 + m_2^2/s_2^2}{(1/s_1^2 + 1/s_2^2)^2}$$

so if $s_2 \gg s_1$, $\text{E}[t] \sim m_1$, $\text{Var}[t] \sim s_1^2 m_1^2$ as required

What is t_0?

Dependence of E[t] on t_0 is **very** weak: if $t_0 \to t_0 + \delta t_0$

$$\delta \mathrm{E}[t] = \delta t_0 \frac{t_0(m_1 - m_2)(s_1^2\sigma_2^2 - \sigma_1^2 s_2^2)}{(\sigma_1^2 + s_1^2 t_0^2 + \sigma_2^2 + s_2^2 t_0^2)^2}$$

Vanishes if

- $m_1 = m_2$
- $s_1 = s_2$ and $\sigma_1 = \sigma_2$
- $s^2 t_0^2 \ll \sigma^2$
- $\sigma^2 \ll s^2 t_0^2$

Rough estimates: for BCDMS & NMC $\delta \mathrm{E}[t] \sim 0.003\, \delta t_0$
for H1NC & CDF-jets $\delta \mathrm{E}[t] \sim 0.02\, \delta t_0$

(a) Nondecoupling MC result ($t_0 = 0$) **very** close to true result
(b) can find t_0 by (very quickly) iterating t to self consistency

Summary

E[t]	Bias $s_1 = s_2$ $\sigma_1 = \sigma_2$		Decoupling $s^2 m^2 \gg \sigma^2$
	1 expt	N expts	2 expts
cov_0	1	1	$\frac{1}{2}(m_1 + m_2)$
m-cov	$1/(1+Nr^2)$	$1-2r^2$	$m_1 m_2 \frac{m_1/s_2^2 + m_2/s_1^2}{m_1^2/s_2^2 + m_2^2/s_1^2}$
n-cov	1	$1-2r^2$	
t_0-cov	1	1	$\frac{m_1/s_1^2 + m_2/s_2^2}{1/s_1^2 + 1/s_2^2}$

Summary

E[t]	Bias $s_1=s_2$	$\sigma_1=\sigma_2$	Decoupling $s^2m^2 \gg \sigma^2$
	1 expt	N expts	2 expts
cov_0	1	1	$\frac{1}{2}(m_1+m_2)$
m-cov	$1/(1+Nr^2)$	$1-2r^2$	$m_1 m_2 \frac{m_1/s_2^2+m_2/s_1^2}{m_1^2/s_2^2+m_2^2/s_1^2}$
n-cov	1	$1-2r^2$	
t_0-cov	1	1	$\frac{m_1/s_1^2+m_2/s_2^2}{1/s_1^2+1/s_2^2}$

NNPDF1.x : $\sim$ 0.1% error

Summary

E[t]	Bias $s_1=s_2$ $\sigma_1=\sigma_2$		Decoupling $s^2m^2 \gg \sigma^2$
	1 expt	N expts	2 expts
cov_0	1	1	$\frac{1}{2}(m_1+m_2)$
m-cov	$1/(1+Nr^2)$	$1-2r^2$	$m_1 m_2 \frac{m_1/s_2^2 + m_2/s_1^2}{m_1^2/s_2^2 + m_2^2/s_1^2}$
n-cov	1	$1-2r^2$	
t_0-cov	1	1	$\frac{m_1/s_1^2 + m_2/s_2^2}{1/s_1^2 + 1/s_2^2}$

D'Agostini Bias ~ 10-20% error

Summary

E[t]	Bias $s_1=s_2$	Bias $\sigma_1=\sigma_2$	Decoupling $s^2m^2 \gg \sigma^2$
	1 expt	N expts	2 expts
cov_0	1	1	$\frac{1}{2}(m_1+m_2)$
m-cov	$1/(1+Nr^2)$	$1-2r^2$	$m_1 m_2 \frac{m_1/s_2^2+m_2/s_1^2}{m_1^2/s_2^2+m_2^2/s_1^2}$
n-cov	1	$1-2r^2$	$m_1 m_2 \frac{m_1/s_2^2+m_2/s_1^2}{m_1^2/s_2^2+m_2^2/s_1^2}$
t_0-cov	1	1	$\frac{m_1/s_1^2+m_2/s_2^2}{1/s_1^2+1/s_2^2}$

Penalty (MSTW): $\sim$ 1% error

Summary

E[t]	Bias $s_1=s_2$ $\sigma_1=\sigma_2$		Decoupling $s^2m^2 \gg \sigma^2$
	1 expt	N expts	2 expts
cov_0	1	1	$\frac{1}{2}(m_1+m_2)$
m-cov	$1/(1+Nr^2)$	$1-2r^2$	$m_1 m_2 \frac{m_1/s_2^2 + m_2/s_1^2}{m_1^2/s_2^2 + m_2^2/s_1^2}$
n-cov	1	$1-2r^2$	
t_0-cov	1	1	$\frac{m_1/s_1^2 + m_2/s_2^2}{1/s_1^2 + 1/s_2^2}$

NNPDF 2.0: Perfect!

Summary

- Nondecoupling MC $\sim 0.1\%$ error
- Hessian Penalty Trick: bias $\sim 1\%$
- t_0-cov : unbiased fits for Monte Carlo (or Hessian)

$$(\mathrm{cov})_{ij} = (\mathrm{cov}_0)_{ij} + s^2 t_0^2$$

Important for balancing DIS and hadronic data