The Gaussian approach to PDFs with errors

S. Alekhin (IHEP, Protvino)

The "Gaussian" approach is based on the the commonly used schemes for inference of the errors from experimental data.

$$y_i = f_i + \sigma_i\mu_i + s_i\lambda$$

y_i – measurements,
f_i – predictions,
σ_i – uncorrelated errors,
s_i – correlated errors,
μ_i, λ – random variables with average=0 and dispersion=1.

Note: μ, λ have **not** to obey Gaussian distribution.

The estimators commonly used

If the errors are uncorrelated ($s = 0$)

$$\chi^2 = \sum_i \frac{(f_i - y_i)^2}{\sigma_i^2}.$$

If the errors are correlated ($s \neq 0$)

- offset of the data by the value of systematic error

$$\chi^2 \to \sum_i \frac{(f_i - y_i - s_i)^2}{\sigma_i^2}$$

(Botje)

- offset with penalty with one parameter for each systematic error

$$\chi^2 = \sum_i \frac{(f_i - y_i - \xi s_i)^2}{\sigma_i^2} + \xi^2$$

(Pascaud-Zomer)

- covariance matrix estimator

$$\chi^2 = \sum_{i,j} (f_i - y_i) E_{ij} (f_j - y_j),$$

$$E_{ij} = (\delta_{ij}\sigma_i\sigma_j + s_i s_j)^{-1}$$

(S.A.)

Theoretical comparison of the estimators

The offset with penalty and covariance matrix estimators are mathematically equivalent.

(Takeuchi 96)
(CTEQ 01)
(Fogli-Lisi-Marrone-Montanino-Palazzo 02)

The offset estimator is inefficient (the error in the fitted parameters are always larger) as compared to the covariance matrix estimator.

(S.A. 00)

Fitting tracks in wire chambers using the Chebyshev norm instead of least squares gives significant reduction of the errors in the parameters of tracks.

(James 82)

Dispersion of the fitted parameter θ in the covariance matrix method is

$$D_{\mathrm{M}}(\hat{\theta}) = \frac{1}{\phi^2}\left[1 + \frac{\rho^2 z^2}{1 + \rho^2(1 - z^2)}\right]$$

$$\rho_i = \frac{s_i}{\sigma_i}, \qquad \phi^i = \frac{f_i'(\theta_0)}{\sigma_i}$$

and in the offset method

$$D_0(\hat{\theta}) = \frac{1}{\phi^2}\left(1 + \rho^2 z^2\right),$$

where z is angle between ρ and ϕ.

If $\rho \gg 1$ ($N \gg 1$ and/or $s \gg \sigma$)

$$D_{\mathrm{M}}^{\mathrm{A}}(\hat{\theta}) \approx \frac{1}{\phi^2(1 - z_1^2)},$$

(saturates)

$$D_0^{\mathrm{A}}(\hat{\theta}) \approx \frac{\rho^2 z^2}{\phi^2}$$

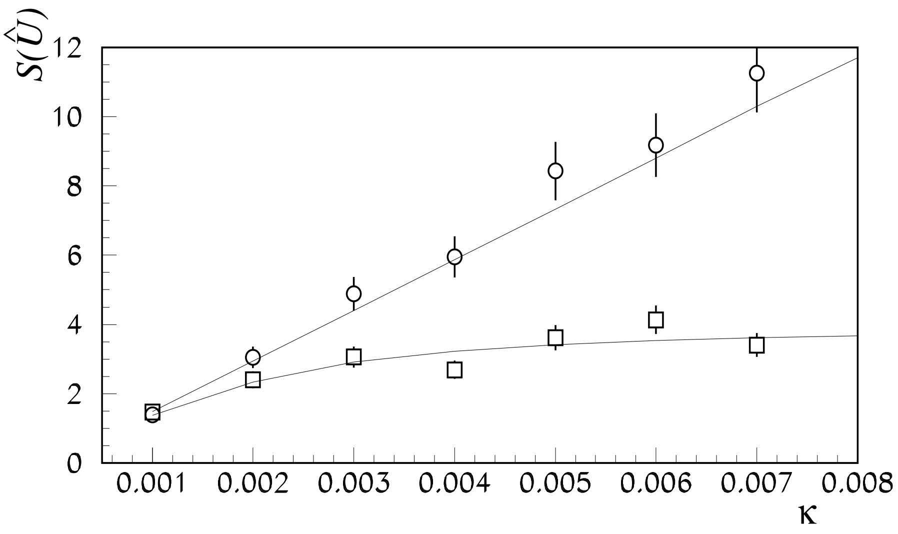
(linearly rises with ρ^2).

Toy model inspired by elastic scattering

$$f_i = U \exp^{(-Vx_i)}, \qquad x_i = 0.1i,$$

where $U = 100, V = 10, i = 1 \ldots 9$. Generate 100 data sets with

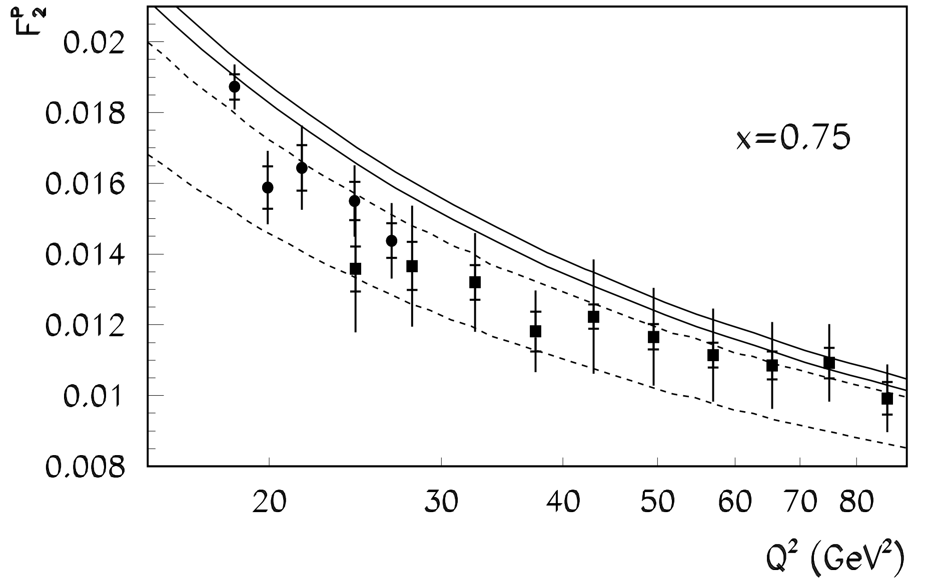
$$\sigma_i = 0.01\sqrt{\frac{U}{f_i}}, \qquad s_i = \frac{\kappa}{x_i}$$

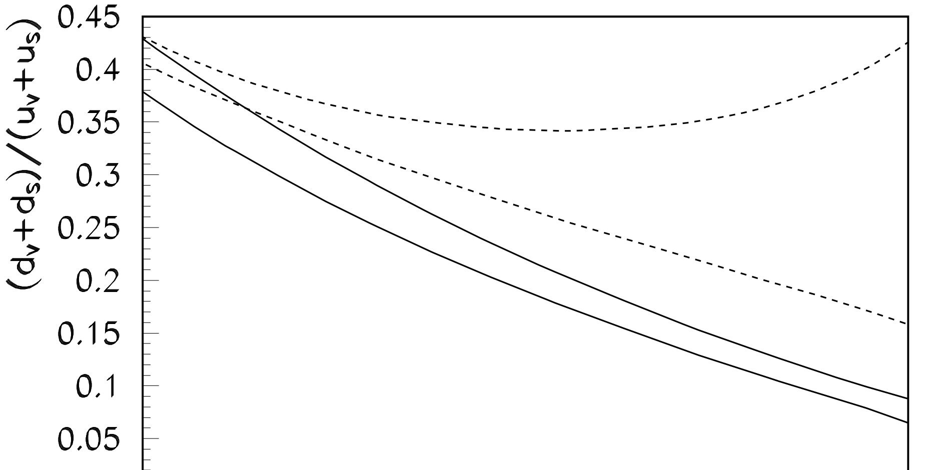


Points: Monte-Carlo calculations for offset method (circles) and covariance method (squares). Curves correspond to analytical formula.

We do observe saturation of the dispersion of covariance matrix estimator

Analysis of real DIS data





Full lines: covariance matrix
Dashed lines: offset

Implementation issues

Problem: The basic difficulty in the implementation is calculation of the derivatives (especially of the second) since the theoretical formula is result of 3-dimensional (or higher) integration. The problem is more difficult if the correlations are non-negligible.

- in the covariance matrix approach the covariance matrix is poorly defined that amplified numerical instabilities in calculation of χ^2.
- in the offset with penalty the numerical instabilities manifest in the inversion of the error matrix of the parameters.

Measures: Improve accuracy at all steps of calculations: accuracy of the evolution code, and calculation of the cross sections from the PDFs with particular attention paid to the scheme used for calculation of the derivatives.

- iterative diagonalization of the error matrix

(CTEQ)

- semianalytical calculation of derivatives

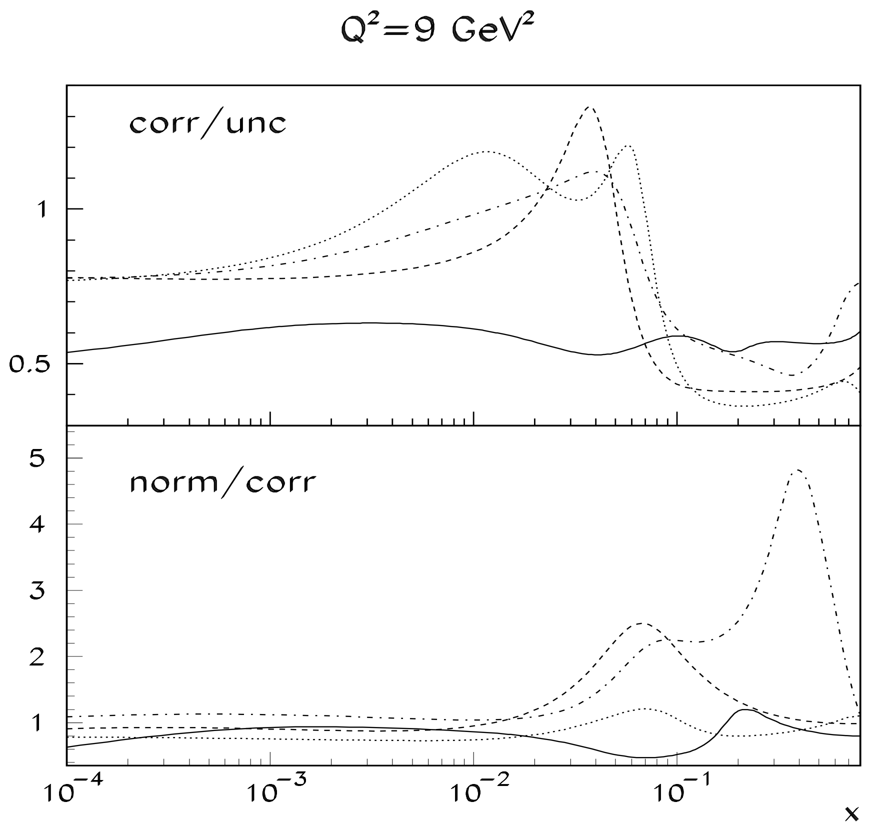
(S.A.)

The first approach seems to be less productive since it did not allow to account for the general normalization errors in the CTEQ's fit, however quantitative comparisons have to be done for the final conclusion

Relative contribution of the correlated, uncorrelated, and normalization error

Global analysis of the DIS data

(S.A. 02)



Solid lines: gluons; dashes: sea; dots: d-quarks; dashed-dots: u-quarks.

Normalization errors give important contribution to the PDFs errors.

Semianalytical scheme

$$G^S = \sum_i \frac{2}{\sigma_i^2} \frac{f_i^+ - f_i^-}{2\delta} \left[f_i(\theta) - y_i \right] =$$

$$= \sum_i \frac{2}{\sigma_i^2} \left[\frac{\partial f_i(\theta)}{\partial \theta} \cdot (f_i(\theta) - y_i) + \frac{\delta^2}{2} \frac{f_i(\theta) - y_i}{6} \cdot \frac{\partial^3 f_i(\theta)}{\partial \theta^3} + O(\delta^4) \right],$$

where $f^{\pm} = f(\theta \pm \delta)$.

Straightforward scheme

$$G^D = \sum_i \frac{2}{\sigma_i^2} \frac{f_i^+ - f_i^-}{2\delta} \left(\frac{f_i^+ + f_i^-}{2} - y_i \right) =$$

$$= G^S + \delta^2 \sum_i \frac{1}{\sigma_i^2} \frac{\partial f_i(\theta)}{\partial \theta} \frac{\partial^2 f_i(\theta)}{\partial \theta^2}.$$

Comparison of semianalytical scheme with the straightforward one

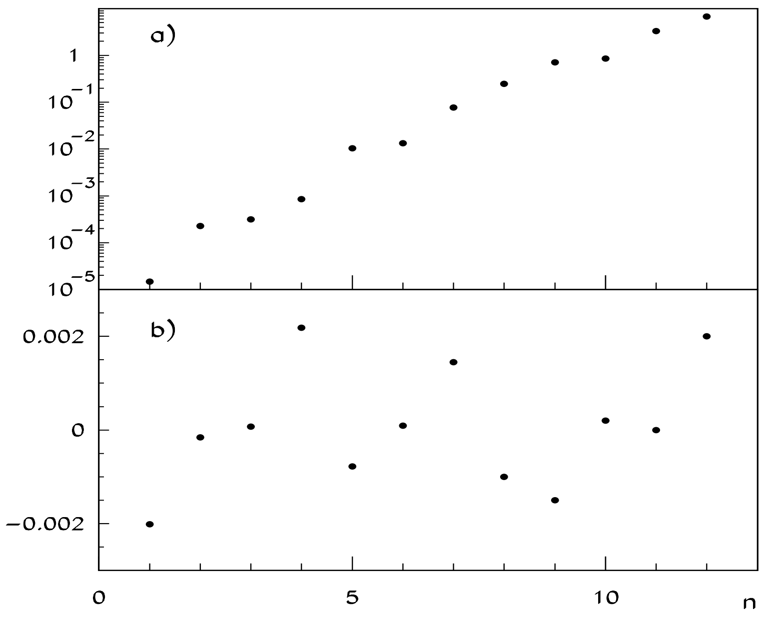
Fit of the phenomenological formula

$$F_2(x, Q^2) = A(x) \left(\frac{\ln(Q^2/\Lambda^2)}{\ln(Q_0^2/\Lambda^2)} \right)^{B(x)} \left(1 + \frac{C(x)}{Q^2} \right),$$

$$A(x) = x^{a_1}(1-x)^{a_2}[a_3 + a_4(1-x)],$$

$$B(x) = b_1 + b_2 x + b_3/(x + b_4), \qquad C(x) = c_1 x + c_2 x^2 + c_3 x^3 + c_4 x^4,$$

to the NMC data on F_2



Upper panel: Eigenvectors of the second derivatives matrix calculated using semi-analytical scheme

Lower panel: Difference between semianalytical and straightforward schemes

Conclusions

Advantages:

The Gaussian approach to the PDFs errors provides economic way to estimate the PDFs' errors. For example, even in the global fit of the PDFs to the DIS data with the iterative account of the target mass and the nuclear corrections (requires up to 5-dimensional integrations) the PDFs errors could be estimated.

Open questions:

- benchmark of the newly developed tools

  (Les Houches 01)

- functional form of the PDFs: possible solution is to use splines

  (Kulagin-Liuti-S.A. 03)

- Is this approach truely Gaussian? (distribution of the PDFs for the non-Gaussian distribution of data)

  (Giele 01)