Error matrix, Lagrange multiplier, data selection, and estimates of contours in global analysis of data

PRD 65, [014011,014012,014013] (2002)

Two issues we want to clarify (related to issues we face in CR studies):

- Can we always rely on Minuit for the estimate of contours, or, equivalently, can we always rely on the error matrix for the contours?
- How to deal with global fits when many data sets are inconsistent with one another?

1) Error/Hessian/Fisher/Information matrix approach: a reminder

2) Error estimate: Fisher matrix vs Lagrange multiplier (to avoid linear approximations)

3) Numerical accuracy and the iterative procedure for the Hessian

4) Global fit and tolerance parameter

Journal Club
27/01/2017

1) Error matrix approach: a reminder

→ Standard approach in cosmology

- The Dark Energy Task Force report (starting p.94)
- The Joint Dark Energy Mission Figure of Merit ScienceWorking Group (starting p.4, similar, but adds cautionary notes on numerical instabilities to watch for)

1.1 Fisher Matrix Overview

Here we review the Fisher Matrix methods used by the DETF. These methods are standard in many fields. First we consider a statistically simple case of a series of measurements with Gaussian error distributions. Suppose we measure the quantity y when the remaining observables have the values x and suppose we put the values of x in bins $b=1,\ldots B$. Suppose in addition that the data should be described by a function f of the bin b and some parameters p and that the expected variance in bin b is σ^2_b, then we can form

$$\chi^2 = \sum_{b=1}^{B} \sum_{i_b} \frac{(f_b(p) - y_{i_b})^2}{\sigma_b^2} \qquad (1.1)$$

where i_b labels the events in bin b. If the parameters p give the true underlying distribution $\bar{p}$, then a Gaussian distribution of data values is:

$$P(y_{i_b}) \propto \exp(-\frac{1}{2}\chi^2) \qquad (1.2)$$

1) Error matrix approach: a reminder

Using Bayes' theorem with uniform prior we have $P(p\,|\,y) \propto P(y\,|\,p)$ so that the likelihood of a parameter estimate can be described as a Gaussian with the same χ^2, now viewed as a function of parameters. If we expand about the true values of the parameters, $p_i = \bar{p}_i + \delta p_i$, and average over realizations of the data,

$$\left\langle \chi^2(p) \right\rangle = \left\langle \chi^2 \right\rangle + \cancel{\left\langle \frac{\partial \chi^2}{\partial p_j} \right\rangle \delta p_j} + \frac{1}{2} \left\langle \frac{\partial \chi^2}{\partial p_j \partial p_k} \right\rangle \delta p_j \delta p_k \ldots \qquad (1.3)$$

where the expectation values are taken at the true values $\bar{p}$. The mean value of the events in bin b is indeed $f_b(\bar{p})$, so the second term vanishes. The distribution of errors in the measured parameters is thus in the limit of high statistics proportional to

$$\exp\left(-\frac{1}{2}\chi^2\right) \propto \exp\left(-\frac{1}{4}\left\langle \frac{\partial \chi^2}{\partial p_j \partial p_k} \right\rangle \delta p_j \delta p_k\right) = \exp\left(-\frac{1}{2} F_{jk} \delta p_j \delta p_k\right) \qquad (1.4)$$

where the Fisher matrix is

[using 1.1] → $$F_{jk} = \sum_b \frac{N_b}{\sigma_b^2} \frac{\partial f_b}{\partial p_j} \frac{\partial f_b}{\partial p_k} \qquad (1.5)$$

and N_b is the average number of events in bin b. From this expression it follows that

$$\left\langle \delta p_j \delta p_k \right\rangle = \left(F^{-1}\right)_{jk} \qquad (1.6)$$

In other words, the covariance matrix is simply the inverse of the Fisher matrix (and vice versa).

1) Error matrix approach: a reminder

More generally, if one can create a probability $P(p_i \mid y_i)$ of the model parameters given a set of observed data, *e.g.* by Bayesian methods, then one can define the Fisher matrix components via

$$F_{ij} = -\left\langle \frac{\partial^2 \ln P}{\partial p_i \partial p_j} \right\rangle$$

and the Cramer-Rao theorem states that any unbiased estimator for the parameters will deliver a covariance matrix on the parameters that is no better than F^{-1}. The Fisher matrix therefore offers a best-case scenario for ones ability to constrain cosmology parameters given a set of observations.

1) Error matrix approach: a reminder

Nice properties of Fisher matrix

Change of variables

If we want to use some other set of parameters q, the new Fisher matrix is simply

$$F'_{lm} = \sum_b \frac{N_b}{\sigma_b^2} \frac{\partial f_b}{\partial q_l} \frac{\partial f_b}{\partial q_m} = \sum_b \frac{N_b}{\sigma_b^2} \frac{\partial p_j}{\partial q_l} \frac{\partial p_k}{\partial q_m} \frac{\partial f_b}{\partial p_j} \frac{\partial f_b}{\partial p_k} = \frac{\partial p_j}{\partial q_l} \frac{\partial p_k}{\partial q_m} F_{jk} \equiv (\mathbf{M})^{\mathrm{T}} \mathbf{F} \mathbf{M} \quad (1.7)$$

using the usual summation convention on j,k.

Priors

A Gaussian prior with width σ can be placed on the ith parameter by adding to the appropriate diagonal element of the Fisher matrix:

$$F_{kl} \rightarrow F_{kl} + \frac{\delta_{ki}\delta_{li}}{\sigma^2} \quad (1.8)$$

which can also be written as

$$\mathbf{F} \rightarrow \mathbf{F} + \mathbf{F}^P \quad (1.9)$$

where in this case $\mathbf{F}^P$ is an extremely simple matrix (with a single non-zero diagonal element).

And also marginalization, combination of data sets...

1) Error matrix approach: a reminder

How to incorporate a "simple" nuisance parameter
(https://arxiv.org/abs/1103.0354v1)

Multiplicative uncertainties provide the simplest example of systematic uncertainties that can be represented by nuisance parameters in profile likelihoods. As an example, let us assume that the integrated luminosity is measured in some auxiliary study, and results in a 2% uncertainty. We would rewrite the likelihood as

$$\mathcal{L} = \prod_{i=1}^{N} \mathcal{P}(n_i|\mu_i)\mathcal{G}(L|\tilde{L}, \sigma_L) \tag{4}$$

for the measured value $\tilde{L} \pm \sigma_L$. The function $\mathcal{G}$ is a normalized Gaussian of mean $\tilde{L}$ and width σ_L, which serves to constrain the value of the new nuisance parameter L to its measured value. Note that it is L and not $\tilde{L}$ that is used to calculate the μ_i. The negative log likelihood is thus

$$-ln\mathcal{L} = \sum_i [-n_i \ln \mu_i + \mu_i] + \boxed{\frac{(L-\tilde{L})^2}{2\sigma_L^2}} \tag{5}$$

and thus the remnant of the Gaussian term can be regarded as a penalty on the negative log likelihood. It is in principle possible to use functions other than Gaussians to constrain the values of the nuisance parameters. In Bayesian terms the constraint functions are simply the prior probability densities of the nuisance parameters.

2) Error estimate: Fisher matrix vs Lagrange multiplier

Pumplin et al., PRD 65, 014011 (2002)

- Standard error propagation

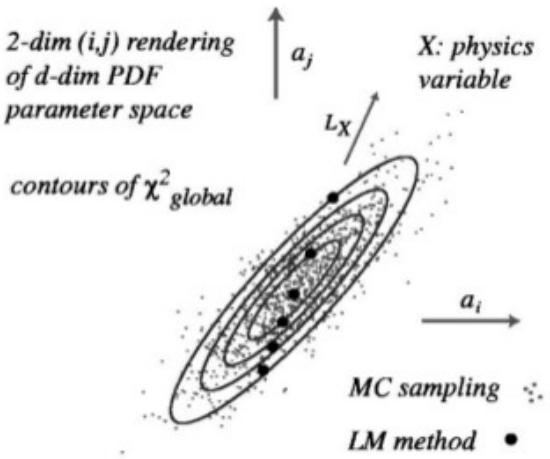
$$\chi^2 = \chi_0^2 + \sum_{i,j} H_{ij} y_i y_j ,$$

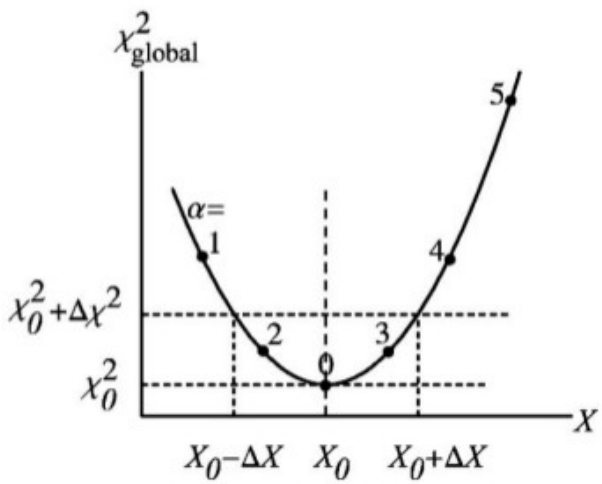
$$H_{ij} = \frac{1}{2}\left(\frac{\partial^2 \chi^2}{\partial y_i \partial y_j}\right)_0 ,$$

$$\longrightarrow \quad (\Delta X)^2 = \Delta\chi^2 \sum_{i,j} \frac{\partial X}{\partial y_i} (H^{-1})_{ij} \frac{\partial X}{\partial y_j} .$$

- Higher order accounting for non-Gaussianity in parameter space (e.g., Sellentin et al., 2014**)**
→ allows banana-shaped contours (beyond simple ellipses)

- Lagrange multiplier approach
→ Find coutour for a constraint $\chi^2 = \chi^2_{\text{min}}$+cst





2) Error estimate: Fisher matrix vs Lagrange multiplier

Wikipedia + Pumplin et al., PRD 65, 014011 (2002)

Consider the two-dimensional problem introduced above

maximize $f(x, y)$
subject to $g(x, y) = 0$.

The method of Lagrange multipliers relies on the intuition that at a maximum, $f(x, y)$ cannot be increasing in the direction of any neighboring point where $g = 0$. If it were, we could walk along $g = 0$ to get higher, meaning that the starting point wasn't actually the maximum.

We can visualize contours of f given by $f(x, y) = d$ for various values of d, and the contour of g given by $g(x, y) = 0$.

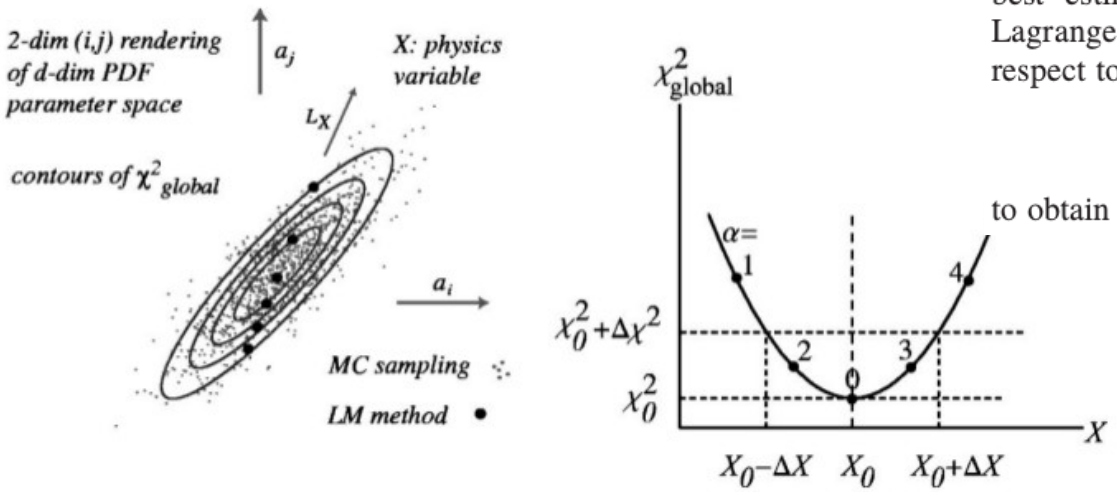
To check the first possibility, notice that since the gradient of a function is perpendicular to the contour lines, the contour lines of f and g are parallel if and only if the gradients of f and g are parallel. Thus we want points (x, y) where $g(x, y) = 0$ and

$$\nabla_{x,y} f = \lambda \nabla_{x,y} g,$$

are the respective gradients. The constant λ is required because although the two gradient vectors are parallel, the magnitudes of the gradient vectors are generally not equal. This constant is called the Lagrange multiplier. (In some conventions λ is precede by a minus sign).

- Lagrange multiplier approach

→ Find coutour for a constraint $\chi^2 = \chi^2_{\text{min}} + \text{cst}$



Let X_0 be the value of X at the χ^2 minimum, which is the best estimate of X. For a fixed value of λ, called the Lagrange multiplier, one performs a new minimization with respect to the fit parameters $\{a_i\}$, this time on the quantity

$$F = \chi^2 + \lambda(X - X_0), \qquad (22)$$

to obtain a pair of values $[\chi^2(\lambda), X(\lambda)]$. (The constant term

2) Error estimate: Fisher matrix vs Lagrange multiplier

$\Delta\chi^2$ for the contours: effect of correlated errors

Stump et al., PRD 65, 014012 (2002) – App. A

Consider an observable m that is measured N times. We shall refer to N measurements of m as one "experiment." Let the true value of m be m_0. The measurements are $m_1, m_2, m_3, \ldots, m_N$. The deviations from the true value are $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_N$, where $\alpha_i = m_i - m_0$. In general, the measurement errors are correlated, so in the Gaussian approximation the probability distribution of the fluctuations is

$$dP = \mathcal{N} \exp\left\{ -\frac{1}{2} \sum_{i,j=1}^{N} \alpha_i C_{ij} \alpha_j \right\} d^N\alpha. \qquad \text{(A1)}$$

tion (A1). For this Gaussian distribution,

$$\langle \alpha_i \alpha_j \rangle = (C^{-1})_{ij}. \qquad \text{(A2)}$$

The mean-square fluctuation E_i of the ith measurement m_i is

$$E_i \equiv \langle \alpha_i^2 \rangle = (C^{-1})_{ii}. \qquad \text{(A3)}$$

To find the best estimate of the value of m from these N measurements, *ignoring the correlations in the measurement errors*, we define a chi-squared function $\chi_u^2(m)$ by

$$\chi_u^2(m) = \sum_{i=1}^{N} \frac{(m_i - m)^2}{E_i}. \qquad \text{(A4)}$$

The value of m that minimizes $\chi_u^2(m)$, call it $\bar{m}$, is then the best estimate of m_0 based on this information. The function

The standard deviation Σ of $\bar{m}$ is the rms fluctuation; that is,

$$\Sigma^2 = \int (\bar{m} - m)^2 dP = \frac{1}{D^2} \sum_{ij} \frac{(C^{-1})_{ij}}{E_i E_j}, \qquad \text{(A7)}$$

where

$$D = \sum_i \frac{1}{E_i}. \qquad \text{(A8)}$$

The question we wish to answer is this: *How much does $\chi_u^2(m)$ increase, when m moves away from the minimum (at $\bar{m}$) by the amount $\pm\Sigma$ that corresponds to one standard deviation of the mean*? The answer to this question is

Example 1. Suppose the measurement errors are uncorrelated; that is,

$$C_{ij} = \delta_{ij}/E_i. \qquad \text{(A11)}$$

Then the standard deviation of the mean $\bar{m}$ is $\Sigma = 1/\sqrt{D}$. Thus for the uncorrelated case, the increase of χ_u^2 corresponding to one standard deviation of the mean is $\Delta\chi_u^2 = 1$. This is the "normal" statistical result: the 1σ range corresponds to an increase of χ^2 by 1.

2) Error estimate: Fisher matrix vs Lagrange multiplier

$\Delta\chi^2$ for the contours: effect of correlated errors

Stump et al., PRD 65, 014012 (2002) – App. A

$$\Delta\chi_u^2 \equiv \chi_u^2(\bar{m}+\Sigma) - \chi_u^2(\bar{m}) = \frac{\sigma^2+Ns^2}{\sigma^2+s^2}. \qquad \text{(A20)}$$

In the limit $s/\sigma \ll 1$, the error correlations in this model become negligible and $\Delta\chi^2$ reduces to the conventional value of 1. But in the limit $s/\sigma \gg 1$, where the error correlations are dominant, $\Delta\chi^2$ approaches N.

Thus for Example 3—a systematic error with 100% correlation between measurements—the increase of χ_u^2 for a standard deviation of $\bar{m}$ is much larger than 1. If s and σ are comparable, then $\Delta\chi_u^2$ is of order N.

The standard deviation Σ of $\bar{m}$ is the rms fluctuation; that is,

$$\Sigma^2 = \int (\bar{m}-m)^2 dP = \frac{1}{D^2}\sum_{ij}\frac{(C^{-1})_{ij}}{E_i E_j}, \qquad \text{(A7)}$$

here

$$D = \sum_i \frac{1}{E_i}. \qquad \text{(A8)}$$

The question we wish to answer is this: *How much does $\chi_u^2(m)$ increase, when m moves away from the minimum (at $\bar{m}$) by the amount $\pm\Sigma$ that corresponds to one standard deviation of the mean*? The answer to this question is

Example 3. For an even more striking example, suppose the N measurements that constitute a single "experiment" are, for $i = 1,2,3,\ldots,N$,

$$m_i = m_0 + y_i + \beta \qquad \text{(A14)}$$

where the y_i are randomly distributed with standard deviation σ, and the measurements are systematically off by the amount β. Suppose that β has a Gaussian distribution with standard deviation s for replications of the "experiment." In

→ *See, e.g., App.B to properly account for correlated systematic errors in the χ^2*

N.B.: if covariance matrix unknown, or if not quadratic, better to add nuisance parameter...

3) Numerical accuracy and iterative procedure for the Hessian

Pumplin et al., PRD 65, [014011,014013] (2002)

$$\chi^2 = \chi_0^2 + \sum_{i,j} H_{ij} y_i y_j ,$$

$$(\Delta X)^2 = \Delta\chi^2 \sum_{i,j} \frac{\partial X}{\partial y_i} (H^{-1})_{ij} \frac{\partial X}{\partial y_j} .$$

Being a symmetric matrix, H_{ij} has a complete set of n orthonormal eigenvectors $V_i^{(k)} \equiv v_{ik}$ with eigenvalues ϵ_k:

$$\sum_i H_{ij} v_{jk} = \epsilon_k v_{ik} , \qquad (4)$$

These eigenvectors provide a natural basis to express arbitrary variations around the minimum; we replace $\{y_i\}$ by a new set of parameters $\{z_i\}$ defined by

$$y_i = \sum_j v_{ij} \sqrt{\frac{1}{\epsilon_j}} z_j .$$

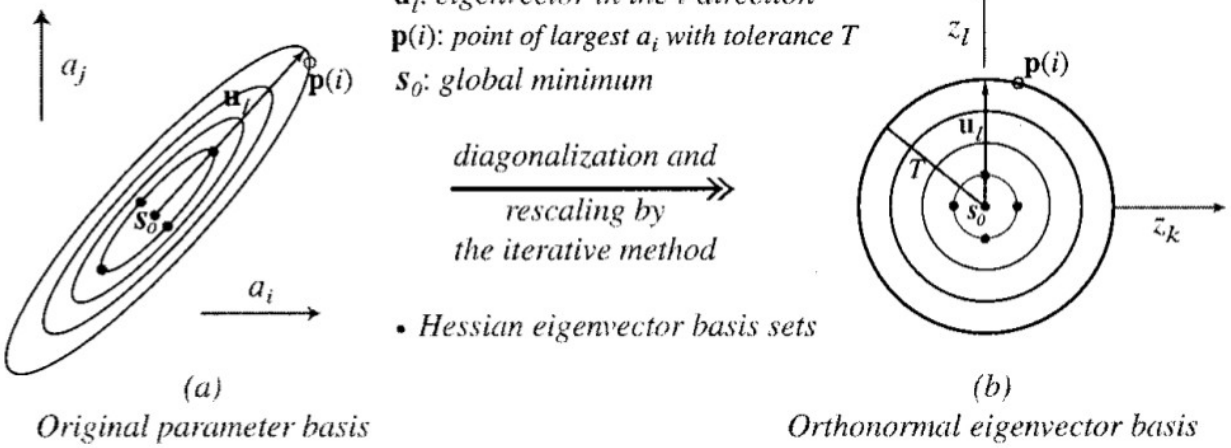
$$\Delta\chi^2 = \chi^2 - \chi_0^2 = \sum_i z_i^2 .$$

$$\Delta X = X - X_0 \cong \sum_i \frac{\partial X}{\partial y_i} y_i = \sum_i X_i z_i ,$$

Goal: avoid numerical inaccuracies/instabilities when calculating the derivatives w.r.t. 'inhomogeneous' variables
→ **may induce a bias of the parameter contours**

N.B: similar in spirit to change of variables (e.g., in cosmology) for better-behaving ones, but going further...

2-dim (i,j) rendition of d-dim (~16) PDF parameter space



3) Numerical accuracy and iterative procedure for the Hessian

Pumplin et al., PRD 65, 014011 (2002)

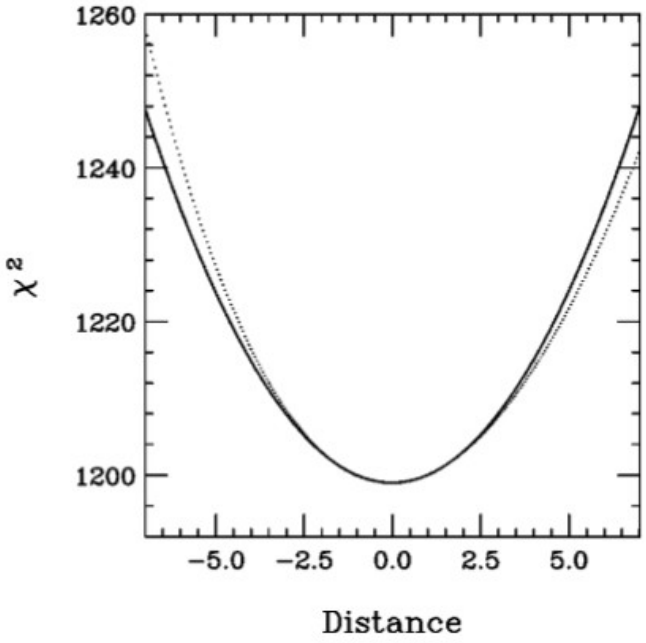


FIG. 1. Variation of χ^2 with distance along a typical direction in parameter space. The dotted curve is the exact χ^2 and the solid curve is the quadratic approximation based on the Hessian. The quadratic form is seen to be a rather good approximation over the range shown.

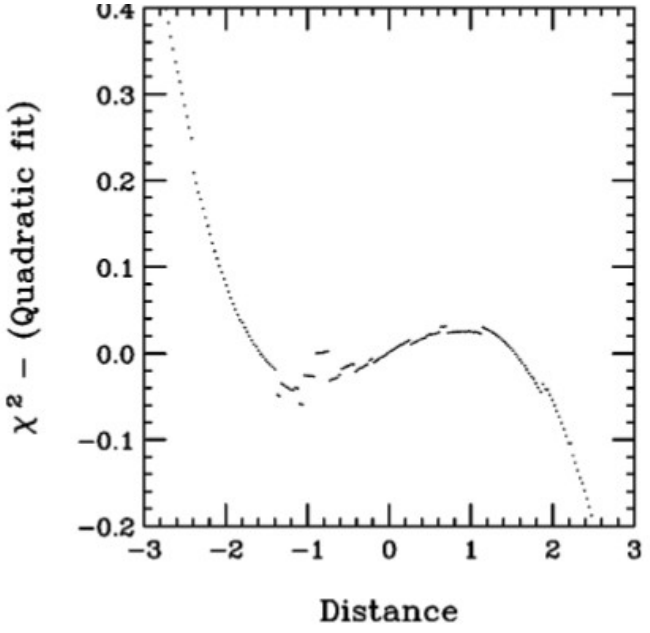


FIG. 2. Difference between χ^2 and its quadratic approximation (2), both of which are shown in Fig. 1. A cubic contribution can be seen, along with a noticeable amount of numerical noise. The fine structure revealed here is small compared to the main variation of χ^2 itself, which rises by 20 over the region shown, as can be seen in Fig. 1.

3) Numerical accuracy and iterative procedure for the Hessian

Pumplin et al., PRD 65, 014011 (2002)

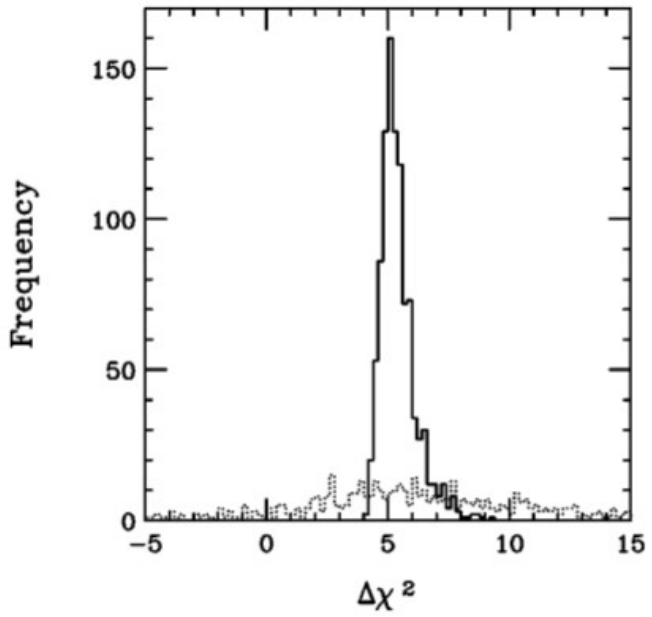


FIG. 3. Frequency distribution of $\Delta\chi^2$ according to the Hessian approximation (2) for displacements in random directions for which the true value is $\Delta\chi^2=5.0$. *Solid histogram*: using Hessian calculated by iterative method of Sec. III; *dotted histogram*: using Hessian calculated by MINUIT.

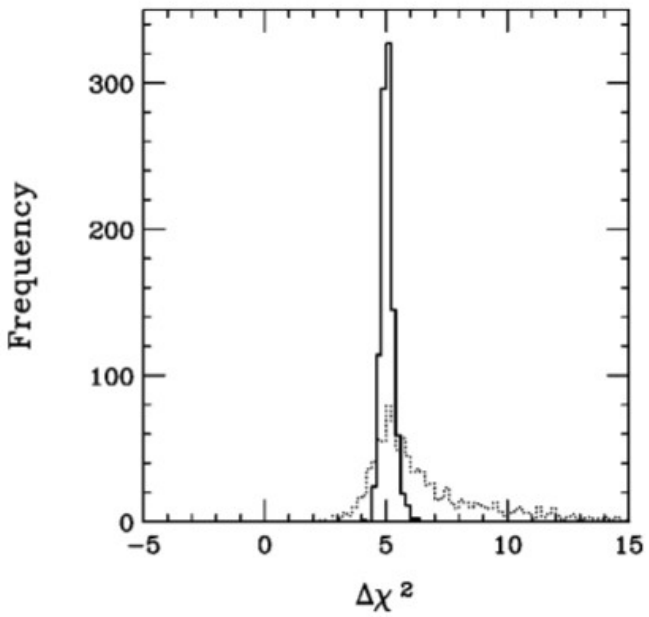


FIG. 4. Same as Fig. 3, except that the displacements are restricted to the parameter subspace spanned by the 10 steepest directions.

3) Numerical accuracy and iterative procedure for the Hessian

Pumplin et al., PRD 65, 014011 (2002)

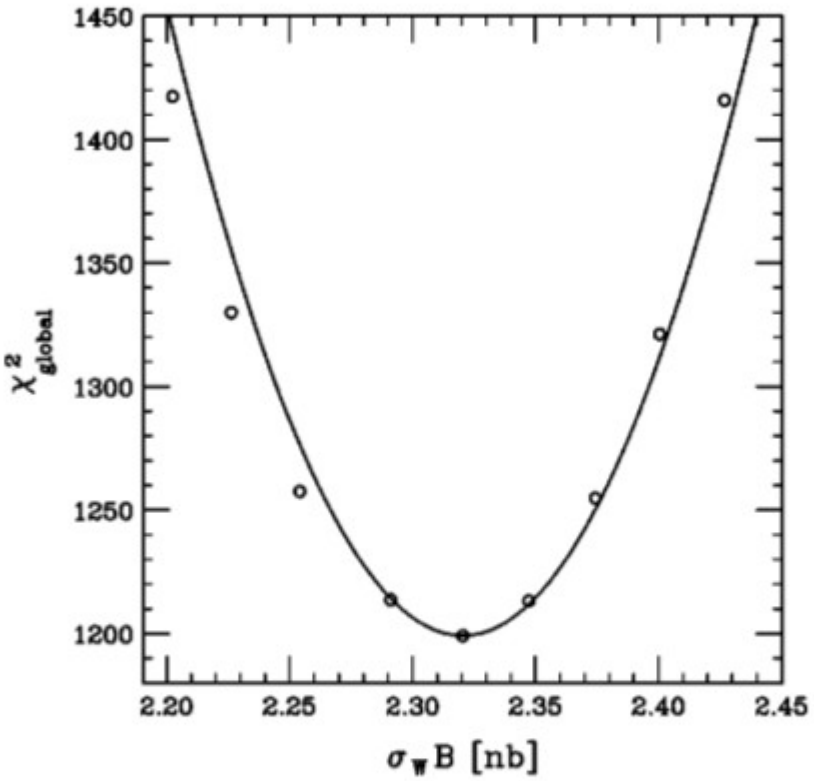


FIG. 5. Minimum χ^2 as a function of the predicted cross section for $W^{\pm}$ production in $p\bar{p}$ collisions. The *parabolic curve* is the prediction of the iteratively improved Hessian method. The *points* are from the Lagrange multiplier method.

4) Global fit and tolerance parameter

Pumplin et al., PRD 65, 014013 (2002) – App. A

[e.g. used for CR data fit/selection in https://arxiv.org/abs/1612.03219]

→ Correlations between point (e.g. spectrum) may be unknown
→ Correlations between different experiments unknown

$$\Delta\chi^2 \neq 1$$

$$\Delta\chi^2 \leq T^2$$

How to determine T?

- Tolerance required by acceptability of the experiments
  - How well best fit agrees with individual datasets: compare with ideal range N_n +/-sqrt($2N_n$)
  - Attribute 'abnormal' χ^2-N_n(sqrt($2N_n$)) to unknown systematic errors/unusual fluctuations
- Tolerance required by mutual compatibility of the experiments
  - If N experiments, fit all combinations of (N-1) experiments
  - max($\Delta\chi^2_{N-1}$) between these combinations indicate that $T^2 >\sim \Delta\chi^2_{N-1}$
- Tolerance calculated from CL of individual experiments
  - Fit individual experiments and calculate errors (e.g., Lagrange multipliers)
  - Alternate fit are considered 'alternative hypotheses'
  - Combine errors and see how uch I requires in terms of $\Delta\chi$ and T

See http://www.desy.de/~blobel/banff.pdf for more on how to correctly deal with systematics, etc.