

## THE “SIEVE” ALGORITHM—SIFTING DATA IN THE REAL WORLD

M. M. BLOCK

*Department of Physics and Astronomy, Northwestern University, Evanston, IL, 60201, USA*  
*E-mail: mblock@northwestern.edu*

Experimental data are rarely, if ever, distributed as a normal (Gaussian) distribution, in real world applications. A large set of data—such as the cross sections for particle scattering as a function of energy contained in the archives of the Particle Data Group<sup>1</sup>—is a compendium of all published data, and hence, unscreened. For many reasons, these data sets have many outliers—points well beyond what is expected from a normal distribution—thus ruling out the use of conventional  $\chi^2$  techniques. We suggest an adaptive algorithm that applies to the data sample a sieve whose mesh is coarse enough to let the background fall through, but fine enough to retain the preponderance of the signal, thus sifting the data. The “Sieve” algorithm gives a robust estimate of the best-fit model parameters in the presence of a noisy background, together with a robust estimate of the model parameter errors, as well as a determination of the goodness-of-fit of the data to the theoretical hypothesis. Computer simulations were carried out to test the algorithm for both its accuracy and stability under varying background conditions.

### 1. Introduction

Our major assumptions about the experimental data are:

- (1) The experimental data can be fitted by a model which successfully describes the data.
- (2) The signal data are Gaussianly distributed, with Gaussian errors.
- (3) That we have “outliers” only, so that the background consists only of points “far away” from the true signal.
- (4) The noise data, *i.e.* the outliers, do not completely swamp the signal data.

### 2. The Adaptive Sieve Algorithm

#### 2.1. Algorithmic steps

We now outline our adaptive Sieve algorithm:

- (1) Make a robust fit of *all* of the data (presumed outliers and all) by minimizing  $\Lambda_0^2$ , the tuned Lorentzian squared, defined as

$$\Lambda_0^2(\boldsymbol{\alpha}; \mathbf{x}) \equiv \sum_{i=1}^N \ln \{1 + 0.179 \Delta\chi_i^2(x_i; \boldsymbol{\alpha})\}, \quad (1)$$

described in detail in Block<sup>2</sup>. The  $M$ -dimensional parameter space of the fit is given by  $\boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_M\}$ ,  $\mathbf{x} = \{x_1, \dots, x_N\}$  represents the abscissas of the  $N$  experimental measurements  $\mathbf{y} = \{y_1, \dots, y_N\}$  that are being fit and  $\Delta\chi_i^2(x_i; \boldsymbol{\alpha}) \equiv \left(\frac{y_i - y(x_i; \boldsymbol{\alpha})}{\sigma_i}\right)^2$ , where  $y(x_i; \boldsymbol{\alpha})$

is the theoretical value at  $x_i$  and  $\sigma_i$  is the experimental error. As discussed in Block<sup>2</sup>, minimizing  $\Lambda_0^2$  gives the same total  $\chi_{\min}^2 \equiv \sum_{i=1}^N \Delta\chi_i^2(x_i; \boldsymbol{\alpha})$  from eq. (1) as that found in a  $\chi^2$  fit, as well as rms widths (errors) for the parameters—for Gaussianly distributed data—that are almost the same as those found in a  $\chi^2$  fit. The quantitative measure of “far away” from the true signal, *i.e.*, point  $i$  is an outlier corresponding to Assumption (3), is the magnitude of its  $\Delta\chi_i^2(x_i; \boldsymbol{\alpha}) = \left(\frac{y_i - y(x_i; \boldsymbol{\alpha})}{\sigma_i}\right)^2$ .

If  $\chi_{\min}^2$  is satisfactory, make a conventional  $\chi^2$  fit to get the errors and you are finished. If  $\chi_{\min}^2$  is not satisfactory, proceed to step 2.

- (2) Using the above robust  $\Lambda_0^2$  fit as the initial estimator for the theoretical curve, evaluate  $\Delta\chi_i^2(x_i; \boldsymbol{\alpha})$ , for the  $N$  experimental points.
- (3) A largest cut,  $\Delta\chi_i^2(x_i; \boldsymbol{\alpha})_{\max}$ , must now be selected. For example, we might start the process with  $\Delta\chi_i^2(x_i; \boldsymbol{\alpha})_{\max} = 9$ . If any of the points have  $\Delta\chi_i^2(x_i; \boldsymbol{\alpha}) > \Delta\chi_i^2(x_i; \boldsymbol{\alpha})_{\max}$ , reject them—they fell through the “Sieve”. The choice of  $\Delta\chi_i^2(x_i; \boldsymbol{\alpha})_{\max}$  is an attempt to pick the largest “Sieve” size (largest  $\Delta\chi_i^2(x_i; \boldsymbol{\alpha})_{\max}$ ) that rejects all of the outliers, while minimizing the number of signal points rejected.
- (4) Next, make a conventional  $\chi^2$  fit to the sifted set—these data points are the ones that have been retained in the “Sieve”. This fit is used to estimate  $\chi_{\min}^2$ . Since the data set has been truncated by eliminating the points with  $\Delta\chi_i^2(x_i; \boldsymbol{\alpha}) > \Delta\chi_i^2(x_i; \boldsymbol{\alpha})_{\max}$ , we must slightly

renormalize the  $\chi_{\min}^2$  found to take this into account, by the factor  $\mathcal{R}$ , whose inverse is shown in Fig. 9a of Block<sup>2</sup>.

If the renormalized  $\chi_{\min}^2$ , *i.e.*,  $\mathcal{R} \times \chi_{\min}^2$  is acceptable—in the *conventional* sense, using the  $\chi^2$  distribution probability function—we consider the fit of the data to the model to be satisfactory and proceed to the next step. If the renormalized  $\chi_{\min}^2$  is not acceptable and  $\Delta\chi_i^2(x_i; \alpha)_{\max}$  is not too small, we pick a smaller  $\Delta\chi_i^2(x_i; \alpha)_{\max}$  and go back to step 3. The smallest value of  $\Delta\chi_i^2(x_i; \alpha)_{\max}$  that makes much sense, in our opinion, is  $\Delta\chi_i^2(x_i; \alpha)_{\max} = 2$ . After all, one of our primary assumptions is that the noise doesn't swamp the signal. If it does, then we must discard the model—we can do nothing further with this model and data set!

- (5) From the  $\chi^2$  fit that was made to the “sifted” data in the preceding step, evaluate the parameters  $\alpha$ . Next, evaluate the  $M \times M$  covariance (squared error) matrix of the parameter space which was found in the  $\chi^2$  fit. We find the new squared error matrix for the  $\Lambda^2$  fit by multiplying the covariance matrix by the square of the factor  $r_{\chi^2}$  (for example<sup>2</sup>,  $r_{\chi^2} \sim 1.02, 1.05, 1.11$  and  $1.14$  for  $\Delta\chi_i^2(x_i; \alpha)_{\max} = 9, 6, 4$  and  $2$ , respectively), shown in Fig. 9b of Block<sup>2</sup>. The values of  $r_{\chi^2} > 1$  reflect the fact that a  $\chi^2$  fit to the *truncated* Gaussian distribution that we obtain—after first making a robust fit—has a rms (root mean square) width which is somewhat greater than the rms width of the  $\chi^2$  fit to the same untruncated distribution. Extensive computer simulations<sup>2</sup> demonstrate that this *robust* method of error estimation yields accurate error estimates and error correlations, even in the presence of large backgrounds.

You are now finished. The initial robust  $\Lambda_0^2$  fit has been used to allow the phenomenologist to find a sifted data set. The subsequent application of a  $\chi^2$  fit to the *sifted set* gives stable estimates of the model parameters  $\alpha$ , as well as a goodness-of-fit of the data to the model when  $\chi_{\min}^2$  is renormalized for the effect of truncation due to the cut  $\Delta\chi_i^2(x_i; \alpha)_{\max}$ . Model parameter errors are found when the covariance (squared error) matrix of the  $\chi^2$  fit is multiplied by the appropriate factor  $(r_{\chi^2})^2$  for the cut  $\Delta\chi_i^2(x_i; \alpha)_{\max}$ .

It is the *combination* of using both  $\Lambda_0^2$  (robust) fitting and  $\chi^2$  fitting techniques on the sifted set that gives the Sieve algorithm its power to make both a robust estimate of the parameters  $\alpha$  as well as a robust estimate of their errors, along with an estimate of the goodness-of-fit.

Using this same sifted data set, you might then try to fit to a *different* theoretical model and find  $\chi_{\min}^2$  for this second model. Now one can compare the probability of each model in a meaningful way, by using the  $\chi^2$  probability distribution function of the numbers of degrees of freedom for each of the models. If the second model had a very unlikely  $\chi_{\min}^2$ , it could now be eliminated. In any event, the model maker would now have an objective comparison of the probabilities of the two models.

### 3. Evaluating the Sieve algorithm

We will give two separate types of examples which illustrate the Sieve algorithm. In the first type, we computer-generated data, normally distributed about

- a constant, along with random noise to provide outliers. The advantage here, of course, is that we know which points are signal and which points are noise.

For our real world example, we took eight types of experimental data for elementary particle scattering from the archives of the Particle Data Group<sup>1</sup>. For all energies above 6 GeV, we took total cross sections and  $\rho$ -values and made a fit to these data. These were all published data points and the entire sample was used in our fit. We then made separate fits to

- $\bar{p}p$  and  $pp$  total cross sections and  $\rho$ -values,
- $\pi^-p$  and  $\pi^+p$  total cross sections  $\sigma$  and  $\rho$ -values,

using eqns. (2) and (3) below.

### 4. Real World data— $\bar{p}p$ and $pp$

We will illustrate the Sieve algorithm by simultaneously fitting all of the published experimental data above  $\sqrt{s} > 6$  GeV for both the total cross sections  $\sigma$  and  $\rho$  values for  $\bar{p}p$  and  $pp$  scattering, as well as for  $\pi^-p$  and  $\pi^+p$  scattering. The  $\rho$  value is the ratio of the real to the imaginary forward scattering amplitude and  $\sqrt{s}$  is the cms energy  $E_{\text{cms}}$ . The data

sets used have been taken from the Web site of the Particle Data Group<sup>1</sup> and have not been modified.

#### 4.1. Testing the Froissart Bound Hypothesis

Testing the hypothesis that the cross sections rise asymptotically as  $\ln^2 s$ , as  $s \rightarrow \infty$ , the four functions  $\sigma^\pm$  and  $\rho^\pm$  that we will *simultaneously* fit for  $\sqrt{s} > 6$  GeV are:

$$\sigma^\pm = c_0 + c_1 \ln\left(\frac{\nu}{m}\right) + c_2 \ln^2\left(\frac{\nu}{m}\right) + \beta_{\mathcal{P}'} \left(\frac{\nu}{m}\right)^{\mu-1} \pm \delta \left(\frac{\nu}{m}\right)^{\alpha-1}, \quad (2)$$

$$\rho^\pm = \frac{1}{\sigma^\pm} \left\{ \frac{\pi}{2} c_1 + c_2 \pi \ln\left(\frac{\nu}{m}\right) - \beta_{\mathcal{P}'} \cot\left(\frac{\pi\mu}{2}\right) \left(\frac{\nu}{m}\right)^{\mu-1} + \frac{4\pi}{\nu} f_+(0) \pm \delta \tan\left(\frac{\pi\alpha}{2}\right) \left(\frac{\nu}{m}\right)^{\alpha-1} \right\}, \quad (3)$$

where the upper sign is for  $pp$  ( $\pi^+p$ ) and the lower sign is for  $\bar{p}p$  ( $\pi^-p$ ) scattering<sup>3</sup>. The laboratory energy is given by  $\nu$  and  $m$  is the proton (pion) mass. The exponents  $\mu$  and  $\alpha$  are real, as are the 6 constants  $c_0$ ,  $c_1$ ,  $c_2$ ,  $\beta_{\mathcal{P}'}$ ,  $\delta$  and the dispersion relation subtraction constant  $f_+(0)$ . We set  $\mu = 0.5$ , appropriate for a Regge-descending trajectory, leaving us 7 parameters. We then require the fit to be anchored by the experimental values of  $\sigma_{\bar{p}p}$  and  $\sigma_{pp}$  ( $\sigma_{\pi^-p}$  and  $\sigma_{\pi^+p}$ ), as well as their slopes,  $\frac{d\sigma^\pm}{d\ln s}$ , at  $\sqrt{s} = 4$  GeV for nucleon scattering and  $\sqrt{s} = 2.6$  GeV for pion scattering. This in turn imposes 4 conditions on the above equations and we thus have three free parameters to fit:  $c_1$ ,  $c_2$  and  $f_+(0)$ .

#### 4.2. $\bar{p}p$ and $pp$ raw scattering data

The raw experimental data for  $\bar{p}p$  and  $pp$  scattering for  $E_{\text{cms}} > 6$  GeV were taken from the Particle Data Group<sup>1</sup>. There are a total of 218 points in these 4 data sets. We fit these 4 data sets *simultaneously* using eq. (2) and eq. (3). Before we applied the Sieve, we obtained  $\chi_{\text{min}}^2 = 1185.6$ , whereas we expected 215. Clearly, either the model doesn't work or there are a substantial number of outliers giving very large  $\Delta\chi_i^2$  contributions. The Sieve technique shows the latter to be the case.

#### 4.3. The results of the Sieve algorithm

We now study the effectiveness and stability of the Sieve. Table 1 contains the fitted results for  $\bar{p}p$  and  $pp$  scattering using 2 different choices of the cut-off,  $\Delta\chi_{i\text{max}}^2 = 4$  and 6. It tabulates the fitted parameters from the  $\chi^2$  fit together with the errors found in the  $\chi^2$  fit, the total  $\chi_{\text{min}}^2$ ,  $\nu$ , the number of degrees of freedom (d.f.) after the data have been sifted by the indicated  $\Delta\chi_i^2$  cut-off and the renormalized  $\chi^2/d.f.$

To get robust errors, the errors quoted in Table 1 for each parameter should be multiplied by the common factor  $r_{\chi^2} = 1.05$ , using the cut  $\Delta = 6$ . See Block<sup>2</sup> for details.

Table 1. The results for a 3-parameter fit to Eqns. 2 and 3. The renormalized  $\chi_{\text{min}}^2/\nu$ , taking into account the effects of the  $\Delta\chi_{i\text{max}}^2$  cut, is given in the row labeled  $\mathcal{R} \times \chi_{\text{min}}^2/\nu$ .

Fitted Parameters	$\Delta\chi_{i\text{max}}^2$	
	4	6
$c_1$ (mb)	$-1.452 \pm 0.066$	$-1.448 \pm 0.066$
$c_2$ (mb)	$0.2828 \pm 0.0061$	$0.2825 \pm 0.0060$
$f(0)$ (mbGeV)	$-0.065 \pm 0.56$	$-0.020 \pm 0.56$
$\chi_{\text{min}}^2$	142.8	182.8
$\nu$ (d.f.)	182	190
$\mathcal{R} \times \chi_{\text{min}}^2/\nu$	1.014	1.040

We note that for  $\Delta\chi_{i\text{max}}^2 = 6$ , the number of retained data points is 193, whereas we started with 218, giving a background of  $\sim 13\%$ . We have rejected 25 outlier points (5  $\sigma_{pp}$ , 5  $\sigma_{\bar{p}p}$ , 15  $\rho_{pp}$  and no  $\rho_{\bar{p}p}$  points) with  $\chi_{\text{min}}^2$  changing from 1185.6 to 182.8. We find  $\chi_{\text{min}}^2/\nu = 0.96$ , which when renormalized for  $\Delta = 6$  becomes  $\mathcal{R} \times \chi_{\text{min}}^2/\nu = 1.04$ , a very likely value with a probability of 0.34.

Obviously, we have cleaned up the sample—we have rejected 25 datum points which had an average  $\Delta\chi_i^2 \sim 40!$  We have demonstrated that the goodness-of-fit of the model is excellent and that we had very large  $\Delta\chi_i^2$  contributions from the outliers that we were able to Sieve out. These outliers, in addition to giving a huge  $\chi_{\text{min}}^2/\nu$ , severely distort the parameters found in a  $\chi^2$  minimization, whereas they were easily handled by a robust fit which minimized  $\Lambda_0^2$ , followed by a  $\chi^2$  fit to the sifted data. Inspection of Table 1 shows that the parameter values  $c_1$ ,  $c_2$  and  $f_+(0)$  effectively do not depend on  $\Delta\chi_{i\text{max}}^2$ , our cut-off choice, having only very small changes compared to the predicted parameter errors. Figure 1 shows the result of the fit of eq. (2) to the sieved

data sample of  $\bar{p}p$  and  $pp$  cross sections. Clearly, this is an excellent fit. Its prediction at the LHC is  $\sigma_{pp} = 107.6 \pm 0.1$  mb.

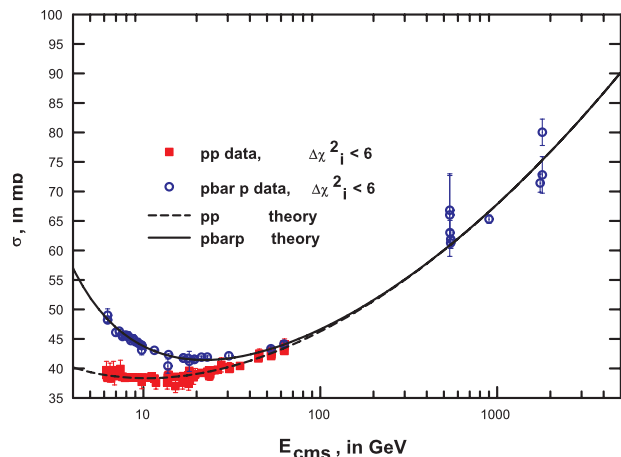


Fig. 1. A plot of  $\sigma_{\bar{p}p}$  and  $\sigma_{pp}$ , in mb vs.  $E_{\text{cms}}$ , the center of mass system energy, in GeV. The data points shown are the result of screening *all* of the cross section points for those points with  $\Delta\chi_i^2 < 6$ . The open circles are  $\sigma_{\bar{p}p}$  and the squares are  $\sigma_{pp}$ . The solid line is the theoretical fit to  $\sigma_{\bar{p}p}$  and the dashed line is the theoretical fit to  $\sigma_{pp}$ .

Due to space limitations, similarly good fits to the  $\rho$  values using eq. (3), as well as  $\sigma_{\pi p}$  and  $\rho_{\pi p}$  fits, are not shown—see ref. 2 for complete details.

## 5. Comments and conclusions

Computer simulations<sup>2</sup> have shown the Sieve algorithm works well in the case of backgrounds in the range of 0 to  $\sim 40\%$ . Extensive computer data were generated about a straight line, as well as about a constant. It also works well for the  $\sim 13\%$  to 19% contamination for the eight real-world data sets taken from the Particle Data Group<sup>1</sup>. However, the Sieve algorithm is clearly inapplicable in the situation where the outliers (noise) swamp the signal. In that case, nothing can be done. See ref. 2 for computer simulation results.

Our particular choice of minimizing the Lorentzian squared in order to extract the robust parameters needed to apply our Sieve technique seems to be a sensible one for both artificial computer-generated noisy distributions, as well as for real-world experimental data. The choice of filtering out all points with  $\Delta\chi_i^2 > \Delta\chi_{i\text{max}}^2$ —where  $\Delta\chi_{i\text{max}}^2$  is as large as possible—is optimal in both minimizing the

loss of good data and maximizing the loss of outliers.

The utilization of the “Sieved” sample with  $\Delta\chi_i^2 < \Delta\chi_{i\text{max}}^2$  allows one to:

- (1) use the *unbiased* parameter values found in a  $\chi^2$  fit to the truncated sample for the cut  $\Delta\chi_i^2(x_i; \alpha)_{\text{max}}$ , even in the presence of considerable background.
- (2) find the renormalized  $\chi_{\text{min}}^2/\nu$ , *i.e.*,  $\mathcal{R} \times \chi_{\text{min}}^2/\nu$ .
- (3) use the renormalized  $\chi_{\text{min}}^2/\nu$  to estimate the goodness-of-fit of the model employing the standard  $\chi^2$  probability distribution function. We thus estimate the probability that the data set fits the model, allowing one to decide whether to accept or reject the model.
- (4) make a robust evaluation of the parameter errors and their correlations, by multiplying the standard covariance matrix  $C$  found in the  $\chi^2$  fit by the appropriate value of  $(r_{\chi^2})^2$  for the cut  $\Delta\chi_{i\text{max}}^2$ .

In conclusion, the “Sieve” algorithm gains its strength from the combination of making first a  $\Lambda_0^2$  fit to get rid of the outliers and then a  $\chi^2$  fit to the sifted data set. By varying the  $\Delta\chi_i^2(x_i; \alpha)_{\text{max}}$  to suit the data set needs, we easily adapt to the different contaminations of outliers that can be present in real-world experimental data samples. Not only do we now have a robust goodness-of-fit estimate, but we also have also a robust estimate of the parameters and, equally important, a *robust* estimate of their errors and correlations. The phenomenologist can now eliminate the use of possible personal bias and guesswork in “cleaning up” a large data set.

## 6. Acknowledgements

I would like to thank Professor Steven Block of Stanford University for valuable criticism and contributions to this manuscript and Professor Louis Lyons of Oxford University for many valuable discussions. Further, I would like to acknowledge the hospitality of the Aspen Center for Physics.

## References

1. K. Hagiwara *et al.* (Particle Data Group), Phys. Rev. D **66**, 010001 (2002).
2. M. M. Block, [arXiv physics/0506010](https://arxiv.org/abs/physics/0506010) (2005).
3. In deriving these equations, we have employed real analytic amplitudes derived using unitarity, analyticity, crossing symmetry, Regge theory and the Froissart bound.