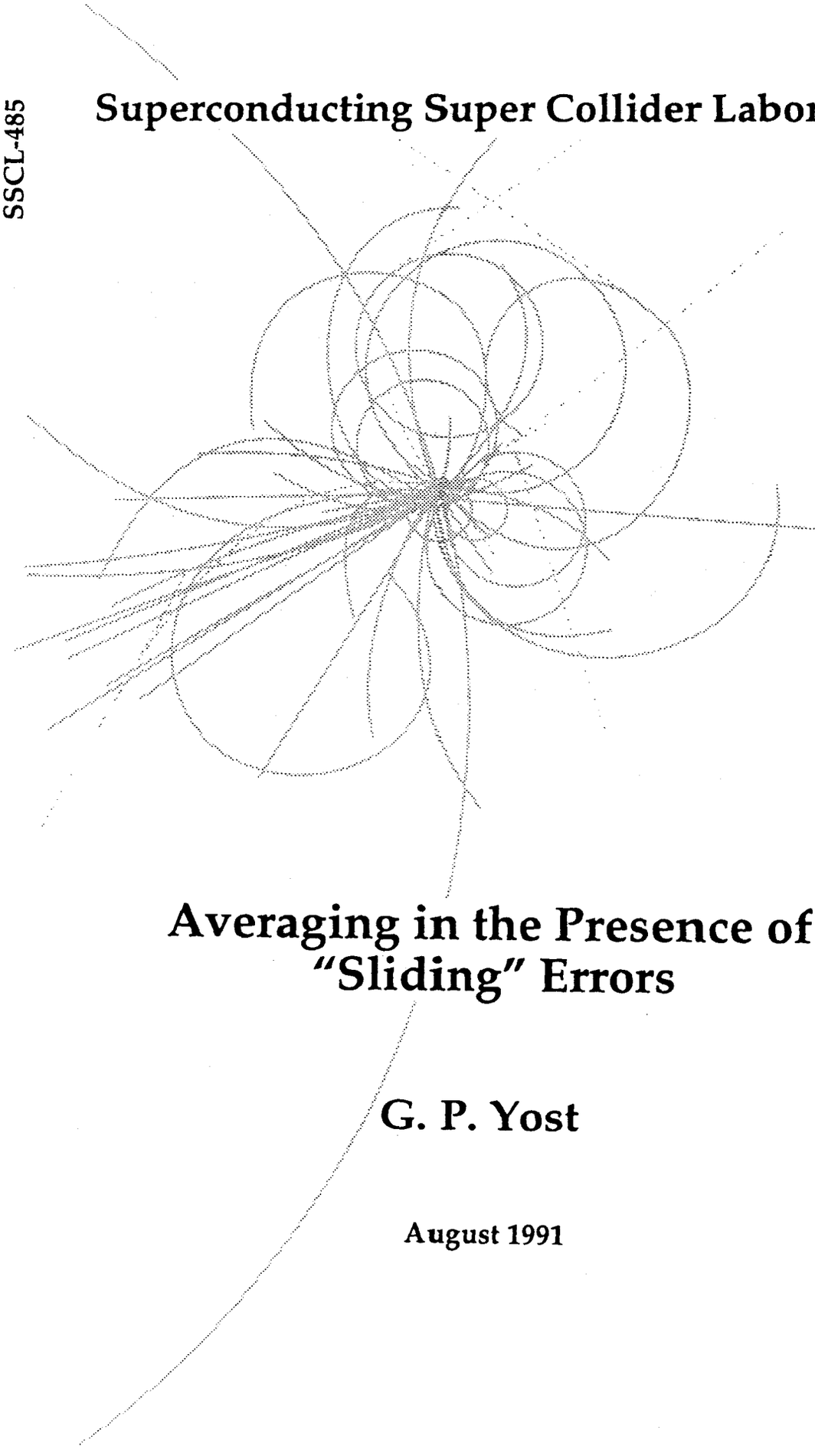
SSCL-485

SSCL-485

Superconducting Super Collider Laboratory



Averaging in the Presence of "Sliding" Errors

G. P. Yost

August 1991

SSCL-485

AVERAGING IN THE PRESENCE OF "SLIDING" ERRORS*

G. P. Yost

Physics Research Division
Superconducting Super Collider Laboratory†
2550 Beckleymeade Ave.
Dallas, TX 75237

and

Department of Physics, University of California
Berkeley, CA 94720

and

Lawrence Berkeley Laboratory
1 Cyclotron Road, Berkeley, CA 94720

August 1991

---

*To be published in Nuclear Instruments and Methods.

†Operated by the Universities Research Association, Inc., for the U.S. Department of Energy under Contract No. DE-AC35-89ER40486.

AVERAGING IN THE PRESENCE OF "SLIDING" ERRORS

G. P. Yost

Physics Research Division
SSC Laboratory†*
2550 Beckleymeade Ave.,
Dallas, TX 75237

and

Department of Physics, University of California,‡
Berkeley, CA 94720

and

Lawrence Berkeley Laboratory††
1 Cyclotron Road, Berkeley, CA 94720

Abstract

In many cases the precision with which an experiment can measure a physical quantity depends on the value of that quantity. Not having access to the true value, experimental groups are forced to assign their errors based on their own measured value. Procedures which attempt to derive an improved estimate of the true value by a suitable average of such measurements usually weight each experiment's measurement according to the reported variance. However, one is in a position to derive improved error estimates for each experiment from the average itself, provided an approximate idea of the functional dependence of the error on the central value is known. Failing to do so can lead to substantial biases. Techniques which avoid these biases without loss of precision are proposed and their performance is analyzed with examples. These techniques are quite general and can bring about an improvement even when the behavior of the errors is not well understood. Perhaps the most important application of the technique is in fitting curves to histograms.

---

† Operated by the Universities Research Association, Inc., for the U.S. Department of Energy under Contract No. DE-AC35-89ER40486.

* Present Address.

‡ Partially supported by the U.S. National Science Foundation under grants NSF PHY-8907526, PHY-8811054, and 85-07153.

†† Operated by the Universities Research Association, Inc., for the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

I. Introduction

Assume that we have n experiments which have measured values v_i which are estimates of some physical quantity which has the true value v_t. We wish to combine these measurements into a best estimate of v_t. For a given technique, an experiment will obtain its value subject to random fluctuations. Neglecting systematic errors, fluctuations may come from two sources: experimental error in measurement, and random variations in the quantity being measured. An important example of the latter is found in measurements of the lifetime of a decaying particle. Each observed decay is a sample from an exponential distribution, in the absence of measurement error. We wish to combine a number of such measurements and estimate the average lifetime. This estimate will be the result v_i of the ith experiment. In the presence of experimental error on each measurement, the distribution from which any given measurement is sampled in this example is a convolution of the exponential and the error distribution.

Unless otherwise stated, the experimental error is not the same for each measurement, even within the same experiment. We will denote the estimated error on the final result v_i by σ_i. In many processes, such as the exponential process, σ_i will depend on v_t. The experimenters will substitute their value v_i in the calculation of σ_i. This may result in a bias when the results of different experiments are combined. This occurs because those values of v_i which result in smaller σ_i will appear to have more precision and exert undue influence on a weighted average.

We propose an averaging technique with superior performance compared with an ordinary weighted average. We shall use three figures of merit:

Bias

$$b = E(\hat{v}) - v_t \quad , \tag{1}$$

where $\hat{v}$ is the improved estimate resulting from combining two or more experiments, and E denotes the expectation operation. We shall always use the "^" to denote a statistic used as an estimator.

Variance

$$V(\hat{v}) = E\{[\hat{v} - E(\hat{v})]^2\}$$

$$= E(\hat{v}^2) - [E(\hat{v})]^2 \quad . \tag{2}$$

Variance denotes the expected square of the spread of $\hat{v}$ which would be found if n independent experiments were combined.

Mean Squared Error (MSE)

$$MSE = E[(\hat{v} - v_t)^2] = V(\hat{v}) + b^2. \tag{3}$$

MSE measures the expected square of the spread of $\hat{v}$ around the true value. MSE allows us to evaluate the merits of cases in which we have decided (or are forced) to accept a certain bias. This has the virtue that it summarizes in one number the other two figures of merit. However, it is often useful to know the breakdown into the bias and variance contributions, and therefore we usually also give this breakdown.

First, assume that the σ_i's do not vary with v. That is, no matter what v_i a given experiment gets, its σ_i is always the same, a characteristic of the experiment. Also assume no bias. Then we can construct a χ^2:

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{v_i - \hat{v}}{\sigma_i} \right)^2 . \tag{4}$$

We shall refer to this as χ^2 even though for non-Gaussian errors it is not distributed as χ^2.

If we select the minimum of this χ^2, we find

$$\hat{v} = \left(\sum v_i/\sigma_i^2 \right) \Big/ \left(\sum 1/\sigma_i^2 \right) . \tag{5}$$

All summations are over experiment index i unless otherwise indicated. This is the ordinary weighted-average result. Since the weight depends on the square of the precision, less precise results, which will tend to be found further from v_t, quickly fade away in importance.

This $\hat{v}$ is at χ^2 minimum only if σ_i is independent of v. If each experiment has Gaussian errors and is itself unbiased, we have a true χ^2 which can be used to test goodness-of-fit. It also gives an unbiased result for $\hat{v}$ since

$$E(\hat{v}) = \left(\frac{\Sigma\, E(v_i)}{\sigma_i^2} \right) \Big/ \left(\Sigma\, 1/\sigma_i^2 \right)$$

$$= \left(v_t\, \Sigma\, 1/\sigma_i^2 \right) \Big/ \left(\Sigma\, 1/\sigma_i^2 \right)$$

$$= v_t \quad . \tag{6}$$

Finally, the variance in $\hat{v}$ is

$$V(\hat{v}) = E(\hat{v}^2) - v_t^2$$

$$= \left[1/\sum 1/\sigma_i^2\right]^2 E\left[\left(\frac{\sum v_i}{\sigma_i^2}\right)^2\right] - v_t^2$$

$$= 1/\sum 1/\sigma_i^2 , \tag{7}$$

where independence has been used to derive

$$E(v_i\, v_j) = v_t^2$$

if $i \neq j$. The measurement with the smallest error is the most influential on both the estimator $\hat{v}$ and its variance. It can also be shown [1] that this $\hat{v}$ has the least possible variance of any unbiased estimator. Therefore we have satisfied our criteria as well as possible and we don't need to look any further.

We can compare this with a simple average of the same data:

$$\hat{v} = \frac{1}{n} \sum v_i \ . \tag{8}$$

Then

$$E(\hat{v}) = v_t$$

$$V(\hat{v}) = \frac{1}{n^2} \sum \sigma_i^2 \ . \tag{9}$$

This estimator is also unbiased, but the error is dominated by the largest σ_i instead of the smallest. The variance in eq. (9) is larger than that in eq. (7) unless all the σ_i's are equal.

II. The Problem

In many cases, the precision of any particular experiment will depend on v_t. For example, assume we have observed in the ith experiment a number n of decays and we want to estimate the true mean lifetime of the particle. Assume the experiment observes decay times with uniform

efficiency out to effectively infinite times. In the simplest case the errors in the individual measurements v_j are negligible (and the background is, also), and the times v_j are sampled from an exponential. We obtain an unbiased estimator with minimum error by taking the average:

$$\hat{v}_i = \frac{1}{n} \sum v_j \tag{10}$$

Then

$$V(\hat{v}_i) = v_t^2/n \quad . \tag{11}$$

Notice in eq. (11) that as v_t increases, so does the error. This reflects the fact that the observations are more spread out in time for larger v_t ; therefore, localizing the mean is harder.

Since no experimenters know v_t, they have to insert v_i for v_t in eq. (11) to estimate their error. Hence measurements with a low v will award themselves a smaller error than those with a high v, even though the true error may be the same in magnitude. If we take a weighted average of such experimental results as in eq. (5) [using $\sigma_i^2 = V(\hat{v}_i)$]we can expect to get a bias favoring measurements which have fluctuated downward.

Consider another common case, a Poisson process. For example, in measuring the decay rate of a particular type of particle into a particular final state, we count the number observed in that decay mode and divide by the total of all decay modes. This frequently will involve corrections for experimental acceptance, so neither numerator nor denominator may be integral. Denote the result for the ith experiment by $v_i = n_i/A_i$. Assuming the error in v_i is dominated by statistical fluctuations in n_i, then we have approximately

$$\sigma_i = \beta_i \sqrt{v_i} \quad . \tag{12}$$

The factor β_i incorporates the acceptance corrections in an average sense, as well as the factor $1/\sqrt{A_i}$: $\sigma_i = \sqrt{n_i} \,/\, A_i = \sqrt{v_i/A_i}$ if the acceptance corrections are unity. Once again, the experimenter has no choice but to insert his own best value for v_i in eq. (12), resulting in a biased estimate for his error unless $v_i = v_t$.

III. Two Approaches

We wish to construct an estimator for v_t by combining two or more experiments. An optimal estimator will have simultaneously the smallest possible bias and variance, and therefore MSE. Therefore we wish to assign more precise results greater weight. If the experiments have error estimates which depend on their measured values v_i , we need to find a way to compensate for this in determining precision. Finally, any technique we use should converge smoothly to eq. (5) in the limit that σ_i is not a function of v_i .

The χ^2 is given by eq. (4). Assume

$$\sigma_i = f_i\,(v_t)\ , \tag{13}$$

where the function f_i is, for the moment, assumed known. Then we calculate

$$\chi^2 = \Sigma \left[\frac{v_i - v}{f_i(v)} \right]^2 \ . \tag{14}$$

In the first approach we will consider, Method I, we assume that $\hat{v}$ is the value of v which minimizes χ^2:

$$\frac{\partial \chi^2}{\partial v} = 2\Sigma \frac{v_i - v}{f_i(v)} \frac{\partial}{\partial v} \left[\frac{v_i - v}{f_i(v)} \right] \tag{15a}$$

and

$$\frac{\partial}{\partial v} \left[\frac{v_i - v}{f_i(v)} \right] = \frac{v - v_i}{f_i^2(v)} \frac{\partial f_i}{\partial v} - \frac{1}{f_i(v)} \ . \tag{15b}$$

The zero of eq. (15a) may be found numerically if necessary.

Method I has appeal in that it formally parallels the standard approach, which looks for the minimum χ^2. However, eq. (14) is only an approximation to the true χ^2 since the denominator is only estimated. Since we vary v, χ^2 is reduced by choosing $\hat{v}$ to overestimate f_i $(\hat{v})$. We will demonstrate this effect below. It results in a bias opposite to the one discussed earlier. However, the bias is considerably reduced and Method I is a significant improvement over the method which neglects the dependence of σ_i on $\hat{v}$.

Method II may be derived either from the exponential or the Poisson example discussed in the previous section. For the exponential, $v_i = \sum_j^{n_i} v_j/n_i$ [eq. (10)] and $f_i = \alpha_i\, v_i$. If n_{tot} events are observed with uniform efficiency (over an unrestricted range) in all experiments combined, the most efficient (smallest variance) unbiased estimator is what would have been obtained if all events had been observed in a single experiment. Thus, $\hat{v} = \sum_j^{n_{tot}} v_j/n_{tot}$, where the sum runs over all events from all experiments.

For the Poisson, $v_i = n_i/A_i$ and $f_i = \beta_i \sqrt{v_i}$. If again n_{tot} events are observed in all experiments, the best estimator is $\hat{v} = n_{tot} / \Sigma A_i$, as if all events were observed in a single experiment.

In both cases, it can be shown that the best estimator is obtained by treating all the data as if it came from a single experiment. This gives the unbiased result with minimum variance. This is also the Maximum Likelihood result [1,2]. These estimators can both be obtained if we take the single step of dropping the term in eq. (15b) which introduced the dependence of the fitted values on the derivative of the errors with respect to those values. Thus, by approximating

$$\frac{\partial}{\partial \hat{v}}\left[\frac{v_i - \hat{v}}{f_i(\hat{v})}\right] \cong -\frac{1}{f_i(\hat{v})} , \quad (15c)$$

we simplify the calculation and improve the estimator at the same time.

To see this in more generality, write

$$f_i(\hat{v}) = \gamma_i \, g(\hat{v}) \quad , \qquad (16)$$

g an arbitrary function. This covers both the exponential and Poisson examples. In the former, $\gamma_i = \alpha_i$ and $g(\hat{v}) = \hat{v}$; in the latter $\gamma_i = 1/\sqrt{A_i} = \beta_i$ and $g(\hat{v}) = \sqrt{\hat{v}}$. We now easily derive

$$\hat{v} = (\Sigma \, v_i / \gamma_i^{\,2})/(\Sigma 1/\gamma_i^2) \quad . \qquad (17)$$

Hence,

$$\hat{v} = (\Sigma \, v_i / \alpha_i^{\,2})/(\Sigma 1/\alpha_i^2) \qquad \text{(Method II)} \qquad (18)$$

and

$$\hat{v} = (\Sigma \, v_i / \beta_i^2)/(\Sigma 1/\beta_i^2) \qquad \text{(Method II)} \qquad (19)$$

for the exponential and Poisson cases, resp. The form of the function $g(\hat{v})$ is irrelevant in the case that the experiment dependence can be factored out of the error as a constant multiplier, eq. (16). This gives us an approach with considerable generality. Likelihood techniques will arrive at the same estimators, but a different calculation is required for each case, since different probability density functions apply. Method II covers a broad spectrum of practical cases with a single formula, eq. (17). This offers hope that we can obtain a reasonable result even when the form of the errors (and, therefore by implication, of the likelihood function) is only poorly known. We address this question in Section V. If acceptance corrections are unity, $\alpha_i = 1/\sqrt{n_i}$ and $\beta_i = 1/\sqrt{A_i}$ in the two examples and Method II yields

$$\hat{v} = \sum_i n_i \, v_i / n_{tot} \qquad (18a)$$

and

$$\hat{v} = \sum_i A_i \; v_i / \sum_i A_i = \Sigma \, n_i / \Sigma \, A_i \quad , \qquad (19a)$$

resp., satisfying the desire to have all the data treated as if they came from a single experiment.

We therefore propose that the solution $\hat{v}$ found by solving

$$\sum_i \frac{\hat{v} - v_i}{f_i^2(\hat{v})} = 0 \tag{20}$$

be used in the general case. The technique is simple to apply and it works optimally for both the uniform-acceptance exponential and Poisson cases. It also works, of course, for the ordinary weighted average when the errors do not depend on the measured value.

To further motivate eq. (15c) we note that the error σ_i should ideally be adjusted using the true value of v, v_t: $\sigma_i = f_i(v_t)$. If we substitute $\hat{v}$ for v_t <u>after</u> evaluating eq. (15b), then we arrive at eq. (15c) directly since $\partial f_i(v_t)/\partial \hat{v} = 0$.

IV. Monte Carlo Studies

We illustrate the problem with a Monte Carlo calculation. For the sake of illustration we assume a truncated Gaussian distribution for the errors of each experiment. The results are scale-invariant, and so without loss of generality we take $v_t = 1$. Initially, we take $\sigma_i = \alpha_i \; v_i$, which is always estimated by the ith experiment as $\alpha_i \; v_i$; α_i is for the moment assumed known without error. In Section V we will discuss questions of incorrect understanding of the form of the errors. This form of $f_i(v)$ corresponds to the exponential case. It has a stronger v dependence than the Poisson case, so we will begin there.

The distribution of 50000 measurements is shown in Fig. 1. Each measurement is chosen at random from a Gaussian distribution with width α, where α is chosen at random from a uniform distribution on the range (0.15, 0.60). This simulates variations among experiments. The distribution is truncated close to zero, leaving 49414 measurements. This truncation results in a 1.6% upward shift in the mean value, but the standard deviation has not changed substantially. The truncation removes unrealistic cases, including negative measurements and those extremely close to zero. We could have left those in the set without damage to the method, but we attempt to focus somewhat more realistically on the measurement of a positive quantity such as a particle lifetime or mass. A 1.6% bias is of little consequence in measurements with an average error of 37.5%, but when measurements are combined we want to be alert to its presence.

We treat each accepted measurement as the outcome of a single experiment. We test our averaging prescriptions by averaging these measurements five at a time.

If we do a simple average without weighting, we obtain the results histogrammed in Fig. 2. The mean is 1.016, so no bias is observed due to the averaging, as expected. The standard deviation, equal to the MSE to three significant digits, is 0.170, which approximately equals $0.375 / \sqrt{5}$, again as expected.

If we do a weighted average according to the prescription of eq. (5) we obtain the results shown in Fig. 3. Here we must take the experimental error estimates at face value, $\sigma_i = \alpha_i \; v_i$. The mean of these weighted average values has an 18% downward bias, due in large part to the tail at very low values. As a comparison, this bias is the same size as the standard deviation of the ordinary average. The peak (most probable value) is around 0.9, and the standard deviation is 0.230. The MSE is 0.085. Clearly, using a weighted average with precision as quoted by the

experimenters is worse than ignoring the relative precisions altogether and simply averaging. This conclusion, obviously, depends on the assumptions made for the distribution of α_i and will not hold in general unless the range of α_i is not large, or one censors data with large α_i (which contribute to the spread in the ordinary average but very little to the weighted average). Note that one must base any decision to censor data on the value of α_i, not $\sigma_i = \alpha_i v_i$.

In Fig. 4 we see the results obtained using Method I. As anticipated, we observe a positive bias of 5%, exceeding the 1.6% built-in bias from the cut. The standard error is 0.214 and the MSE is 0.048. The net result is therefore better than the weighted average but still worse than the ordinary average.

In Fig. 5 we see the results from Method II. The mean is 1.007, consistent with 1.016, and the standard deviation is 0.145. The MSE is 0.020, much lower than for the other methods. This method obtains the best results of any we have considered.

For the Poisson case, we use the same error distribution as above. This will suffice to illustrate the point. The typical error is large, corresponding to a small number of events or to the presence of systematic errors. We now assume eq. (12) describes each experiment's error estimate. That is, β_i is given the value $\sigma_i / \sqrt{v_i}$, where σ_i is, as stated earlier, generated uniformly on the range (0.15, 0.60).

The results are summarized as follows. The regular weighted average (Fig. 6) gets an 8-9% bias, a width of 17.7%, and an MSE of 0.038. These results are much better than for the previous case because the v dependence of the errors is only $\sqrt{v}$ instead of v.

We shall now proceed directly to Method II, which appears to be superior to Method I. In the present example, Method II yields results identical to those of the previous example (Fig. 5). This illustrates the fact that the form of g ($\hat{v}$) [eq. (16)] is irrelevant to the results of Method II. Therefore, Method II again yields an unbiased estimator with the smallest variance.

Two more examples will help clarify the comparison of these methods. We wish to average two measurements of the same quantity. In the first example, the measurements yield 100 ± 20 and 50 ± 10, and we assume $\sigma_i = \alpha_i \hat{v}$. Method I gives $\hat{v} = 83.3 \pm 12.42$ (error estimation will be discussed below), with $\chi^2 = 5.56$. Since $\alpha_i = \sigma_i / \hat{v}_i$ we have $\alpha_1 = \alpha_2$, so both measurements have the same real precision. Therefore, a simple average gives the best answer, and Method II agrees. The regular weighted average gives $\hat{v} = 60.0 \pm 8.9$ with a $\chi^2 = 11.81$. This is more than a standard deviation from the optimal answer.

In the second example, the measurements yield 100 ± 10 and 64 ± 8. We assume now $\sigma_i = \beta_i \sqrt{\hat{v}}$ so $\beta_1 = \beta_2 = 1$. Method I yields $\hat{v} = 83.95 \pm 6.4$ with $\chi^2 = 7.81$. Method II yields $\hat{v} = 82.0 \pm 6.4$ with $\chi^2 = 7.90$. Here there is not much real difference. Since the β's are identical, in this Poisson case the true precisions are the same, and again a simple average seems most reasonable. The regular weighted average gives $\hat{v} = 78.0 \pm 6.2$ with a χ^2 of 8.72.

Method II is clearly superior to the others, and we shall concentrate our attention there for what follows.

V. Robustness

In practical cases we can't be sure how accurately the form of f ($\hat{v}$) is known. For example, there may be systematic errors whose $\hat{v}$dependence is at best poorly understood. We wish to understand the performance of Method II under the circumstances that f($\hat{v}$) is incorrectly parametrized. We will adopt certain forms of f ($\hat{v}$) and test the performance using simplified, incorrect, assumptions. We will demonstrate that this approach is robust in the sense that one obtains an improvement over an ordinary weighted average for a wide class of $\hat{v}$dependencies.

We assume a $\hat{v}$ dependence which is slower than linear at low measured values and steeper at high measured values (Fig. 7). The experimenters measure values on the abcissa of Fig. 7 and assign the errors σ_i shown on the ordinate. Two sample linear dependencies are sketched for comparison. The value of v_t is 1.0, as before. All experiments have true error σ_i (v_t) = 0.3, for the illustration of Fig. 7. To study Method II we now inject a further note of realism: we allow the true precision of each experiment to vary uniformly on the range (0.15, 0.60), as in the preceding section, while maintaining the $\hat{v}$ dependence shown in Fig. 7. The experimental results are generated randomly according to a Gaussian of the resultant width with mean 1.0.

The results of the ordinary weighted average are shown in Fig. 8. There is a 16% bias. The width is 17%, and the MSE is 0.052. Fig. 9 shows the result of Method II. We have deliberately assumed $f_i(\hat{v}) = \alpha_i \hat{v}$, as in the straight-line examples shown in Fig. 7, rather than the correct, complex, $\hat{v}$- dependence of the errors. The mean is 0.992, revealing a negligible bias. The width is 13% and the MSE is 0.016, substantially better. Thus, Method II is robust (in the sense above) for this problem.

To study the case of an error contribution which is independent of $\hat{v}$, we next take

$$\sigma_i = f_i(\hat{v}) = \alpha_i \hat{v} + \varepsilon_i \ . \tag{21}$$

The "flat" component ε_i is chosen uniformly from the range (0.,|2[v_i - 1]|). Such errors, for the example $\alpha_i = 0.3$ for all i, are illustrated in Fig. 10. To study the performance of the estimators we choose $\alpha_i = 0.3$ for all i, but multiply the resultant σ_i by a uniform random number chosen from the range (0.15, 0.60). Because $\varepsilon_i = 0$ at $v_i = v_t = 1$, the true precision of each experiment is 0.30 multiplied by this random number.

This type of error dependence may be encountered in experiments with systematic errors. It also resembles lifetime measurement problems where, for a single event,

$$V(\tau) = \tau^2 + \sigma^2 \ . \tag{22}$$

In this case, τ is the true value of the lifetime (v_t) and σ is the width of a Gaussian representing an error in measuring the individual lifetime. Such an error often appears in impact parameter approaches. The lifetime problem has also been considered by Lyons, Martin, and Saxon [3].

The weighted average result is shown in Fig. 11. The bias is 12%, the width is 16.5%, and the MSE is 0.040. The results are better than on Fig. 8 because of the presence of the flat component; the regular weighted average would work perfectly if this component were dominant.

The Method II results are shown in Fig. 12. We again assume, incorrectly, $f_i(\hat{v}) = \alpha_i' \hat{v}$, where $\alpha_i' = \sigma_i / v_i$. The bias is 2.6%. The width is 14% and the MSE = 0.018, a considerable improvement over Fig. 11. In fact, this is almost ideal; using the correct $f_i(\hat{v})$ gives zero bias, the same width, and improves the MSE by only 0.0004.

VI. Estimating Confidence Intervals (Standard Errors)

In doing least-squares fitting approximate errors can be computed from the values at which

$$\chi^2 = \chi^2_{min} + 1 \quad . \tag{23}$$

This is a well-known technique [4, see also 1,2] which, in a linear least-squares fit gives the same variance as the relation

$$V(\hat{v}) = 2 \Big/ \frac{\partial^2 \chi^2}{\partial v^2}\Big|_{\hat{v}} \quad . \tag{24}$$

In our case neither relation can be, strictly speaking, applied since the fit is not linear and the χ^2, eq. (14), is not at minimum for the solution for Method II.

However, it is reasonable to fix the experimental errors σ_i (v) at their values at solution, σ_i $(\hat{v})$, and proceed to apply either eq. (23) or (24). With the σ_i thus fixed the χ^2 is at minimum. This converges in the appropriate limit to the correct weighted-average answer, eq. (7), because σ_i is constant. Thus, we arrive at the point that we can use well-understood least-squares error estimation techniques.

To understand the performance of these errors we will look at "pulls", defined by

$$P = (\hat{v} - v_t) / \sigma(\hat{v}) \quad . \tag{25}$$

Here, σ $(\hat{v})$ is the estimated error in the solution $\hat{v}$ $\left[\sigma(\hat{v}) = \sqrt{V(\hat{v})}\right]$. One should be aware that pulls are often alternately defined as $P_i = (\hat{v} - v_i) / \sigma_i$ $(\hat{v})$, but in our Monte Carlo calculation we can take advantage of our knowledge of v_t, since we want to understand the performance of the approaches. The pull expresses the actual error in terms of estimated standard deviations. Ideally, the pulls should be Gaussian—distributed with mean zero and width one.

For the case $\sigma_i = f_i(v) = \alpha_i \hat{v}$ (Figs. 3, 4, and 5) the pulls for the regular weighted average are shown in Fig. 13. The average is a full 2 $\hat{\sigma}$ low, where $\hat{\sigma}$ is that value we estimate for the error in $\hat{v}$. The problem with the weighted average is clearly seen: one assumes the resolution to improve with decreasing v_i.

The pulls for Method II are shown in Fig. 14. We have estimated the errors as described above. The bias is -0.1 and the width is close to 1.0, although there is still a tail to low values.

The probability of exceeding n $\hat{\sigma}$ is determined by integrating the | pull | . The results for Method II are given in Table I for various cases with n = 1, 2, and 3. The results are within a few percent of the probabilities achieved for perfectly Gaussian errors. This is true even for cases in which the incorrect form of σ_i (v) has been assumed. One standard deviation estimates are in all cases conservative. That is, the interval $\hat{v} \pm \sigma_i$ covers the true value somewhat more than the 68.3% we anticipate for an accurate Gaussian. Thus, adoption of these techniques does not lead to a false and sometimes dangerous under-estimation of uncertainties.

It again is clear that Method II is a substantial improvement over ordinary weighted averaging. The error in the solution as estimated herein is reasonable. The small deviations from Gaussian behavior in the pulls must be considered in the context of real experimental measurements which will have error contributions from systematic effects, from non-Gaussian statistical errors, and so on.

VII. The SCALE Factor

The Particle Data Group [5] has adopted a practice of increasing the estimated error in their average in the case that the final χ^2 exceeds n–1. Here n is, of course, the number of measurements being averaged. The assumption being made is that the data are partly inconsistent. A conservative approach protects science from drawing restrictive conclusions which may not be justified by the data. The SCALE is a multiplicative factor equal to $[\chi^2 / (n-1)]^{1/2}$, which causes the final χ^2 to equal exactly (n-1). Again, it is only applied when $\chi^2 > (n-1)$.

Of course, some fraction of the time $\chi^2 > (n-1)$ even in a perfect world: for n = 5, for example, this occurs 40% of the time. Therefore, for n = 5 the use of a SCALE factor would result in errors which would be too large 40% of the time if all experiments understood their errors perfectly, and all errors were Gaussian. Retrospective studies by the Particle Data Group [5] suggest, however, that this increase of the errors has been justified.

For the linear errors case ($\sigma_i = \alpha_i v_i$), the ordinary weighted average gives a SCALE factor distribution as shown in Fig. 15. As always in our discussions, n = 5 is assumed. Using propagation of errors to second order, we expect a mean of about 0.94 and a standard deviation of about $1/\sqrt{8} \cong 0.35$. One may also compare with the "chi" distribution, i.e., the distribution of $\sqrt{\chi 2}$ (see ref. [1], chapter 5). Based on Fig. 15, 42% of the time SCALE > 1.0, compared with 40% expected.

These are the only cases in which SCALE would be used by the Particle Data Group. About 26% of the time SCALE is > 1.25, compared with 18% expected, ideally. Thus, the SCALE tends to be overly conservative in this case. Since the regular weighted average has been shown to have numerous problems, this extra conservatism can only be beneficial.

For Method II the corresponding SCALE distribution is given in Fig. 16. The fraction above 1.0 is about 35%, and above 1.25 about 15%. These are not conservative. However, the assumption underlying application of the SCALE is that the experiments are partly discrepant. The Monte Carlo data we have used are not more discrepant than expected from statistics. Therefore, it is not clear that one should worry. We are, if anything, not enlarging the errors as often as we expect for these non-discrepant examples.

The results for other cases we have analyzed are summarized in Table II. It should be emphasized that none of these cases involve discrepant measurements. In all cases shown the SCALE factor is not as large as expected from a chi distribution. This does not compromise the ability of a SCALE to recognize and compensate for discrepant experiments. Rather, it means that the errors are not as prone to be unnecessarily enlarged as for ordinary weighted averaging.

VIII. Some Real Examples

The ratio of branching ratios for the $\Lambda_c^+ \, p\bar{K}^* (892)^0/pK^-\pi^+$ has been measured at (1) 0.42 ± 0.24 and (2) 0.18 ± 0.10 by two experiments. This gives a weighted average of 0.216 ± 0.092. Let us assume this is approximately a Poisson case, with an error in the denominator which is negligible compared with the error in the numerator.

Therefore we assume $\sigma_i = f_i(\hat{\nu}) = \beta_i \sqrt{\hat{\nu}}$. Substituting ν_i in place of ν for each measurement separately we find $\beta_1 = 0.24/\sqrt{0.42} = 0.37$ and $\beta_2 = 0.24$. The experiments have much more nearly the same precision than the original errors, taken literally, would suggest. Then, from eq.(20),

$$\hat{\nu} = [0.42/0.37^2 + 0.18/0.24^2]\,[1/0.37^2 + 1/0.24^2]^{-1} = 0.251 \pm 0.101.$$

The error has increased over the weighted average error of 0.092 because the central value has increased. We now get a scale factor of 1.1, instead of <1 as before.

Also for the Λ_c^+, the ratio of ratios of $\Delta\,(1232)^{++}\,K^-/pK^-\pi^+$ is measured at (1) 0.40 ± 0.17 and (2) 0.17 ± 0.07. The ordinary weighted average yields 0.203 ± 0.081, incorporating a scale factor of 1.3. Applying Method II, we first find $\beta_1 = 0.269$ and $\beta_2 = 0.170$. Therefore $\hat{\nu} = 0.236 \pm 0.112$, *including* the scale factor, which is now 1.6.

Both of these scale factors have gone up, a consequence of the fact that there are only two measurements and the smaller measurement had the greater true precision in both cases. If we switched the β's we would have seen a decrease in scale.

IX. Fitting to Histograms

The χ^2 of eq. (4) and therefore, these techniques, can be applied to the fitting of curves to histograms. Each bin of a histogram is filled by a Poisson process with an expectation value $\hat{\nu}_i$ which depends upon the bin. For n bins we can re-cast eq. (4):

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{\nu_i - \hat{\nu}_i}{\sigma_i} \right)^2 . \tag{26}$$

We assume $\hat{\nu}_i$ is known as a function of a set of fit parameters. After we write $\sigma_i = f_i(\hat{\nu}_i)$ we can proceed to a solution in the same manner as in Method II. For any parameter θ_k,

$$\frac{\partial \chi^2}{\partial \theta_k} = -2 \sum \frac{v_i - \hat{v}_i}{f_i^2(\hat{v}_i)} \frac{\partial \hat{v}_i}{\partial \theta_k} \quad , \tag{27}$$

and therefore the solution satisfies

$$\sum_{i=1}^{n} \frac{v_i}{f_i^2(\hat{v}_i)} \frac{\partial \hat{v}_i}{\partial \theta_k} = \sum_{i=1}^{n} \frac{v_i}{f_i^2(\hat{v}_i)} \frac{\partial \hat{v}_i}{\partial \theta_k} \quad . \tag{28}$$

The solution vector can be found numerically using standard techniques. Assuming $f_i(\hat{v}_i) = \beta_i \sqrt{\hat{v}_i}$ (where $\beta_i = 1$ in the case of equally-weighted events in the histogram bins), eq. (28) simplifies a little:

$$\sum_{i=1}^{n} \frac{v_i}{\beta_i^2 \hat{v}_i} \frac{\partial \hat{v}_i}{\partial \theta_k} = \sum_{i=1}^{n} \frac{1}{\beta_i^2} \frac{\partial \hat{v}_i}{\partial \theta_k} \quad . \tag{29}$$

Standard errors may be found as in Section VI. If the fit is highly non-linear or the correlations are large, it is probably best to estimate these from the contour defined by eq. (23), with $f_i(\hat{v}_i)$ fixed at solution.

X. Conclusions

Method II out-performs the regular weighted-average technique currently in widespread use in cases where the experimental error varies with the value measured. It also outperforms Method I, although by a smaller margin. Method I would doubtless be acceptable most of the time. However, since Method II is simpler to use and gives superior results, it is to be preferred.

These results can be applied to other problems, including the fitting of curves to histograms, which are filled by Poisson processes. Many people use fixed errors based on the number of events observed. This gives a clear bias, since bins fluctuating downward are awarded smaller errors. Method I is the usual alternative. What is shown here is that this will give a small bias in the other direction. Method II should be best, although, as we have shown, for the Poisson cases these two methods do not seriously differ most of the time.

TABLE I
Probability of being within $n\,\hat{\sigma}$, Method II
(Percent)

Case	n=1	2	3
Perfect Gaussian	68.3	95.5	99.7
Linear errs; correctly used	70.6	92.7	97.0
Complex errs; assumed linear	75.9	93.8	98.3
$\sqrt{n}$ errs; correctly used	69.5	94.9	98.9
Lin. + flat errs; assumed linear	76.1	96.1	99.1
Lin. + flat errs; correctly used	73.2	94.6	98.6

XI. Acknowledgements

We gratefully acknowledge fruitful discussions with Ken McFarlane.

TABLE II
SCALE factor behavior

Case	mean	width	% >1.0	% >1.25
Ideal world (approx.)	0.94	0.35	40	18
Wtd. Avg. (lin. errs)	1.06	0.58	42	26
Linear errs; correctly used	0.91	0.35	35	15
Complex errs; assumed linear	0.81	0.28	20	5
$\sqrt{n}$ errs; correctly used	0.91	0.33	35	13
Lin. + flat errs; assumed lin.	0.79	0.28	19	5
Lin. + flat errs; correctly used	0.85	0.30	26	8

REFERENCES

[1] A. G. Frodesen, 0. Skjeggestad, and H. Toefte, *Probability and Statistics in Particle Physics* (Universitetsforlaget, Toeyen, Norway, 1979).

[2] L. Lyons, *Statistics for Nuclear and Particle Physicists* (Cambridge University Press, 1986).

[3] L. Lyons, A.J. Martin, and D.H. Saxon, *On the Determination of the B-lifetime by combining the results of different experiments*, Phys. Rev. D, to be published.

[4] W. T. Eadie, D. Drijard, F. E. James, M. Roos, and B. Sadoulet, *Statistical Methods in Experimental Physics* (North Holland, Amsterdam, and London, 1971).

[5] Particle Data Group, *The Review of Particle Properties,* Phys. Lett. **B239** (1990).

Figure Captions

1. Raw measurements for 50,000 Monte Carlo experiments; lower tail truncated leaving 49414 events. Mean = 1.016, standard deviation = 0.375.

2. Results of simple averaging of the Monte Carlo measurements, 5 at a time. Desired mean = 1.016. Observed mean = 1.016, standard deviation = 0.170.

3. Results of regular weighted average of same samples as in Fig. 2. Assumes each measurement assigned an error proportional to v, the measured value. Observed mean = 0.822, standard deviation = 0.230.

4. Results of Method I, minimum χ^2 averaging of same samples as in Figs. 2 and 3. Assumes each measurement assigned an error proportional to v, the measured value. Observed mean = 1.046, standard deviation = 0.214.

5. Results of Method II, approximate minimum χ^2 averaging of samples in Figs. 2, 3, and 4. Assumes each measurement assigned an error proportional to v, the measured value. Observed mean = 1.007, standard deviation = 0.145.

6. Results of regular weighted average of same samples as in Fig. 2. Assumes each measurement assigned an error proportional to $\sqrt{v}$, where v is the measured value. Observed mean = 0.913, standard deviation = 0.177.

7. Spectrum of complex error assignments. Each measurement yields a value from the abcissa and assigns error read on ordinate. Subsequent Method II averaging will use linear dependence illustrated by straight lines.

8. Results of regular weighted averaging, assuming error assignments based on Fig. 7 with additional uniform "smearing" to simulate varying true precisions. Observed mean = 0.845, standard deviation = 0.171.

9. Results of Method II averaging assuming same errors assigned as in Fig. 8. Errors are incorrectly taken linear in the course of averaging to test robustness of the technique. Observed mean = 0.992, standard deviation = 0.132.

10. "Flat plus linear" errors assigned by experimenters to values measured as on abcissa. To be used in a robustness test for Method II.

11. Results of regular weighted averaging in the presence of errors assigned as in Fig. 10 with an additional uniform "smearing" to simulate varying true precisions. Observed mean = 0.884, standard deviation = 0.166.

12. Results of Method II averaging in the presence of errors assigned as in Fig. 11. Errors are incorrectly taken linear in the course of averaging to test robustness of the technique. Observed mean = 1.026, standard deviation = 0.137.

13. "Pulls" for regular weighted averaging technique in the presence of linear error assignments. The pull is the number of standard deviations the result is in error, based on the calculated standard deviation at the resultant value. Observed mean = -2.191, standard deviation = 3.686.

14. "Pulls" for Method II averaging in the same case as Fig. 13. Observed mean = -0.091, standard deviation = 1.053.

15. Scale factor for the regular weighted averaging technique in the same case as Fig. 13. Observed mean = 1.060, standard deviation = 0.579.

16. Scale factor for Method II averaging in the same case as Fig. 13. Observed mean = 0.914, standard deviation = 0.346.

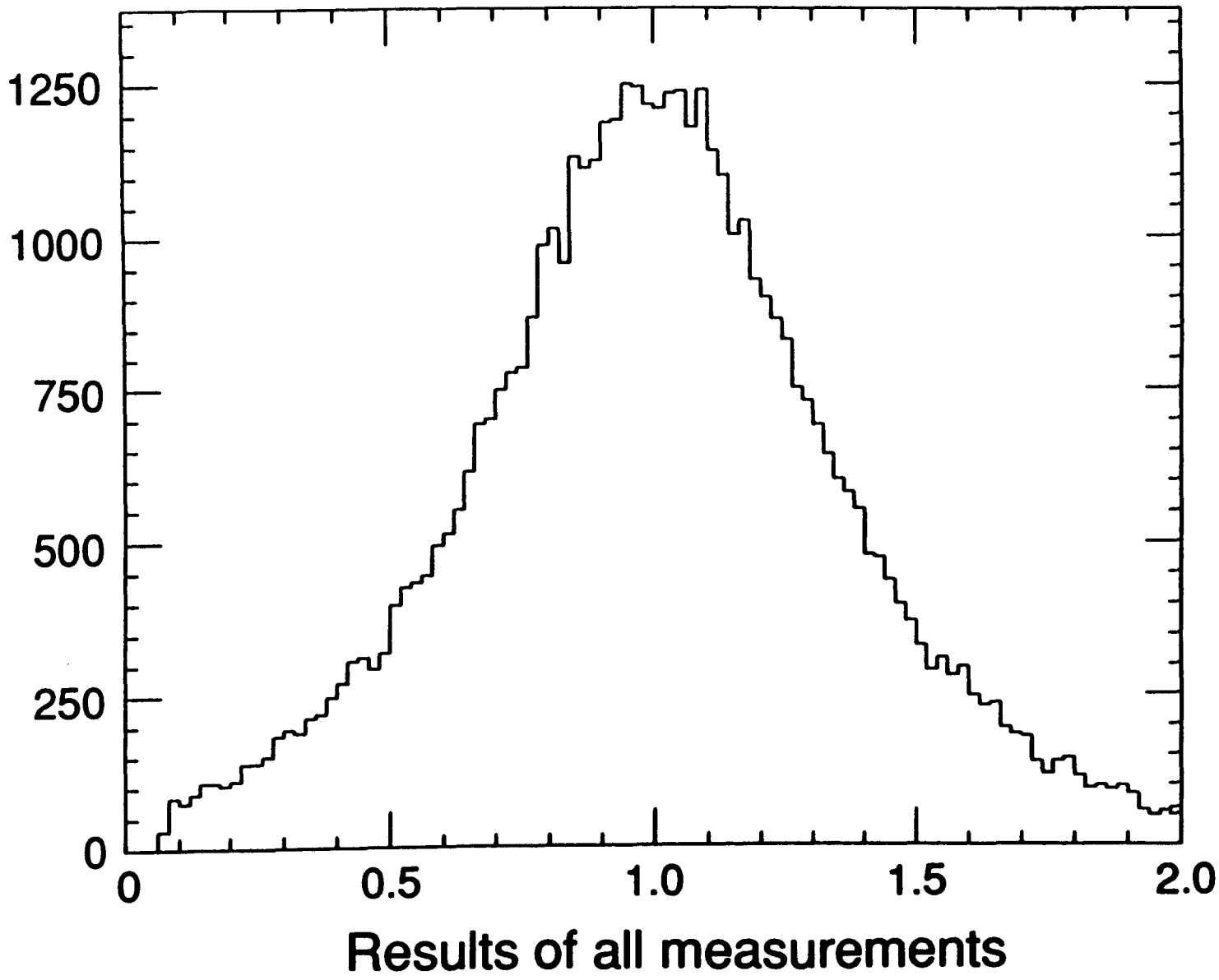


Fig. 1

XBL 898-7712

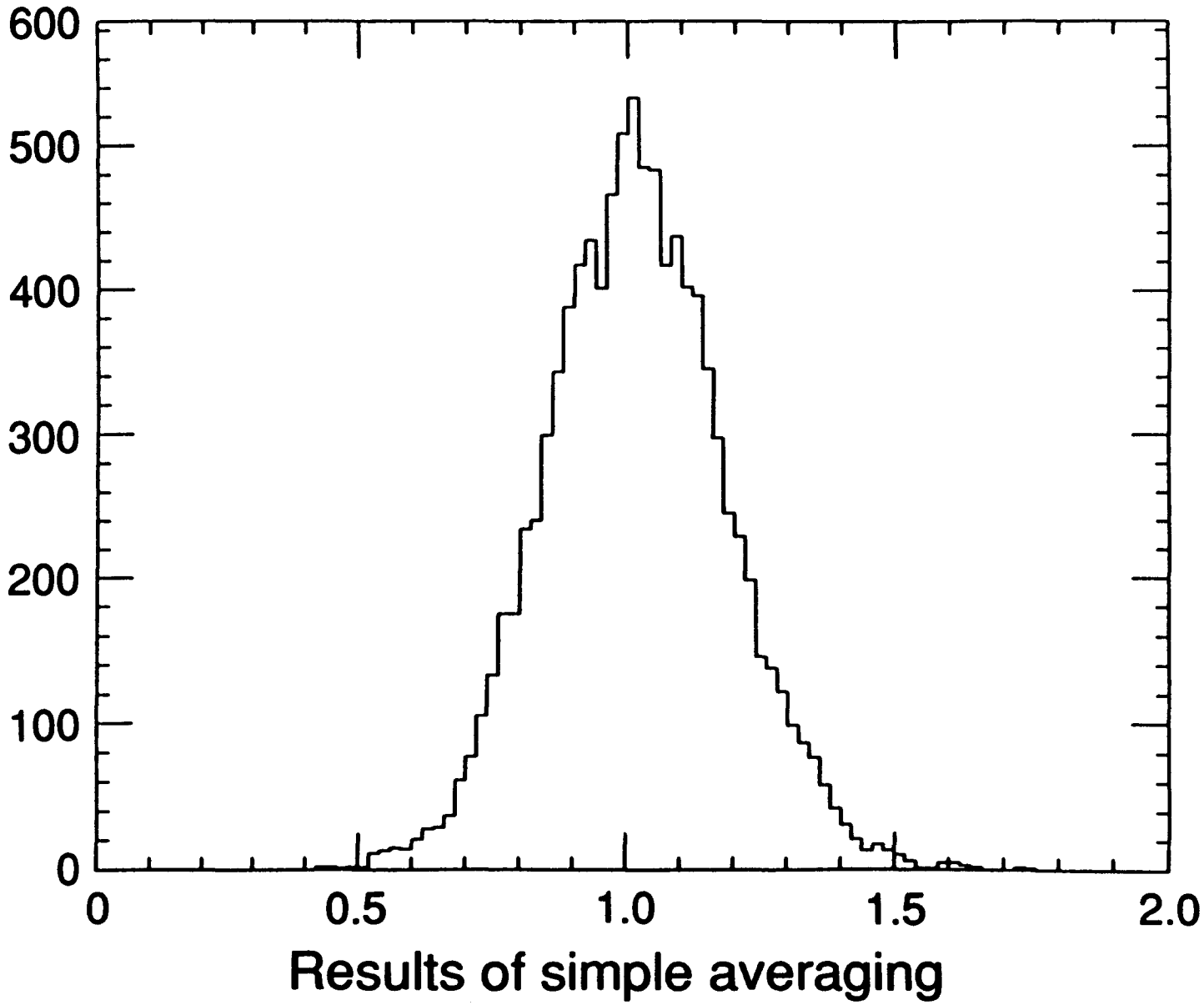
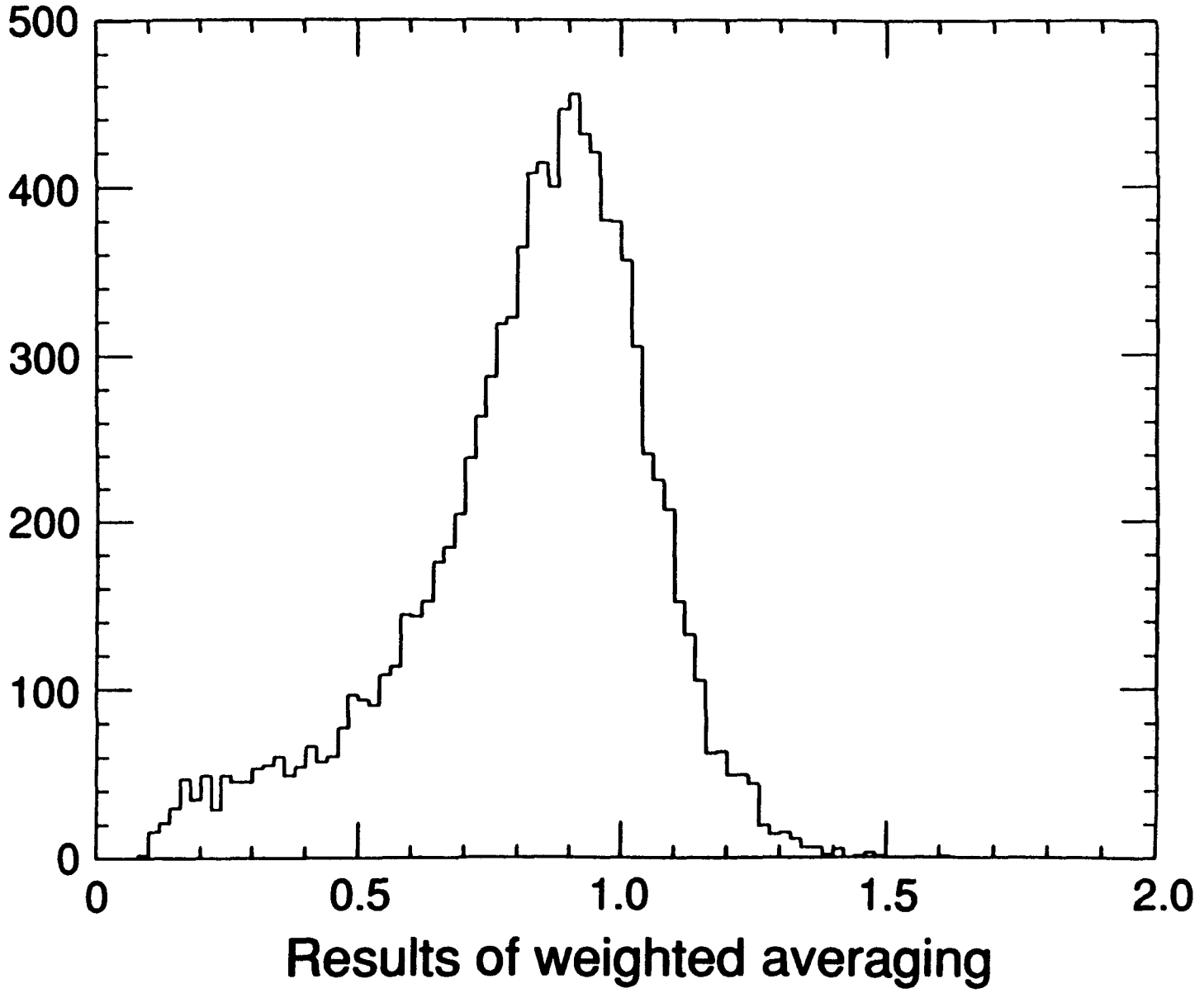


Fig. 2

XBL 898-7711



XBL 898-7710

Fig. 3

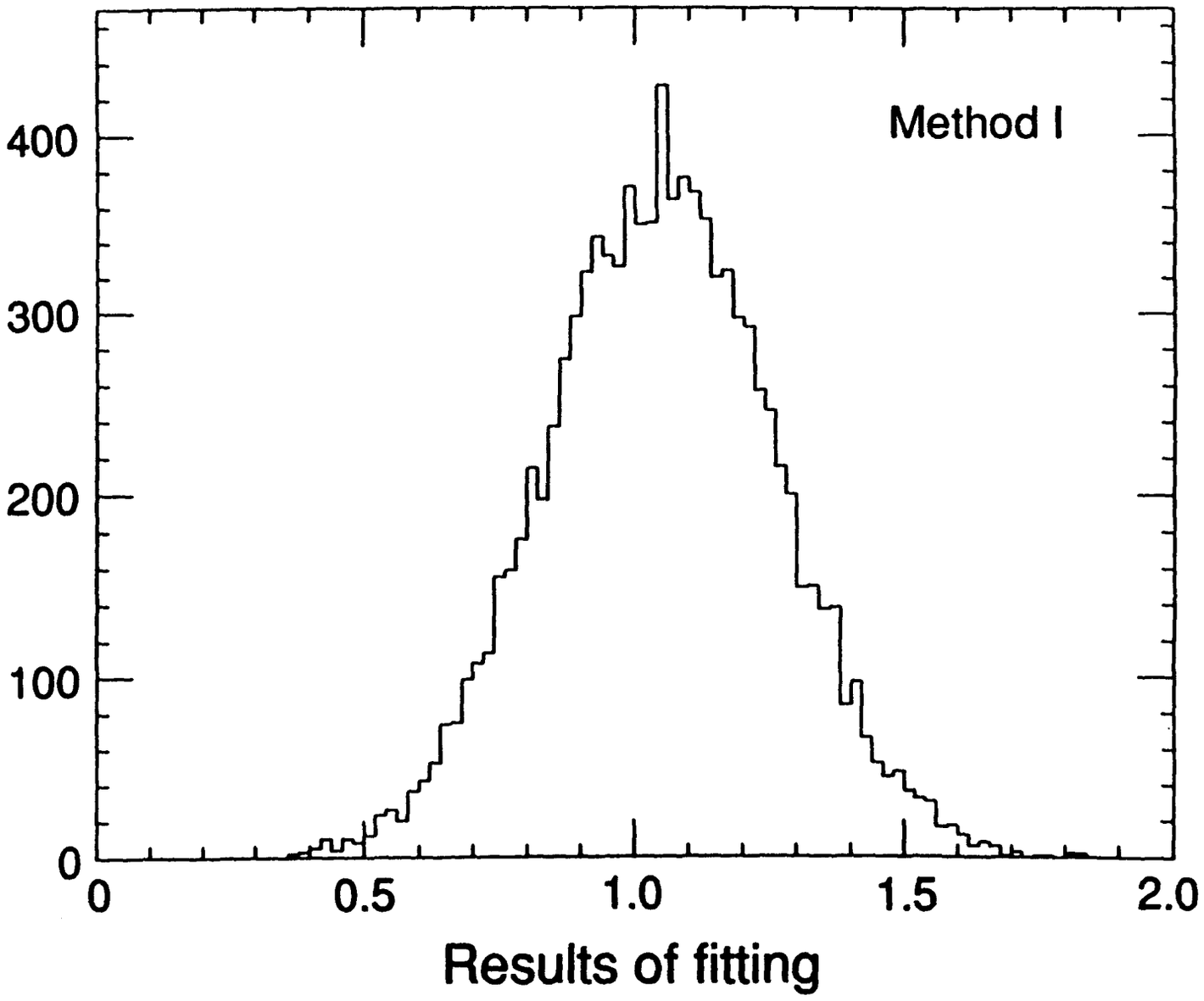


Fig. 4

XBL 898-7709

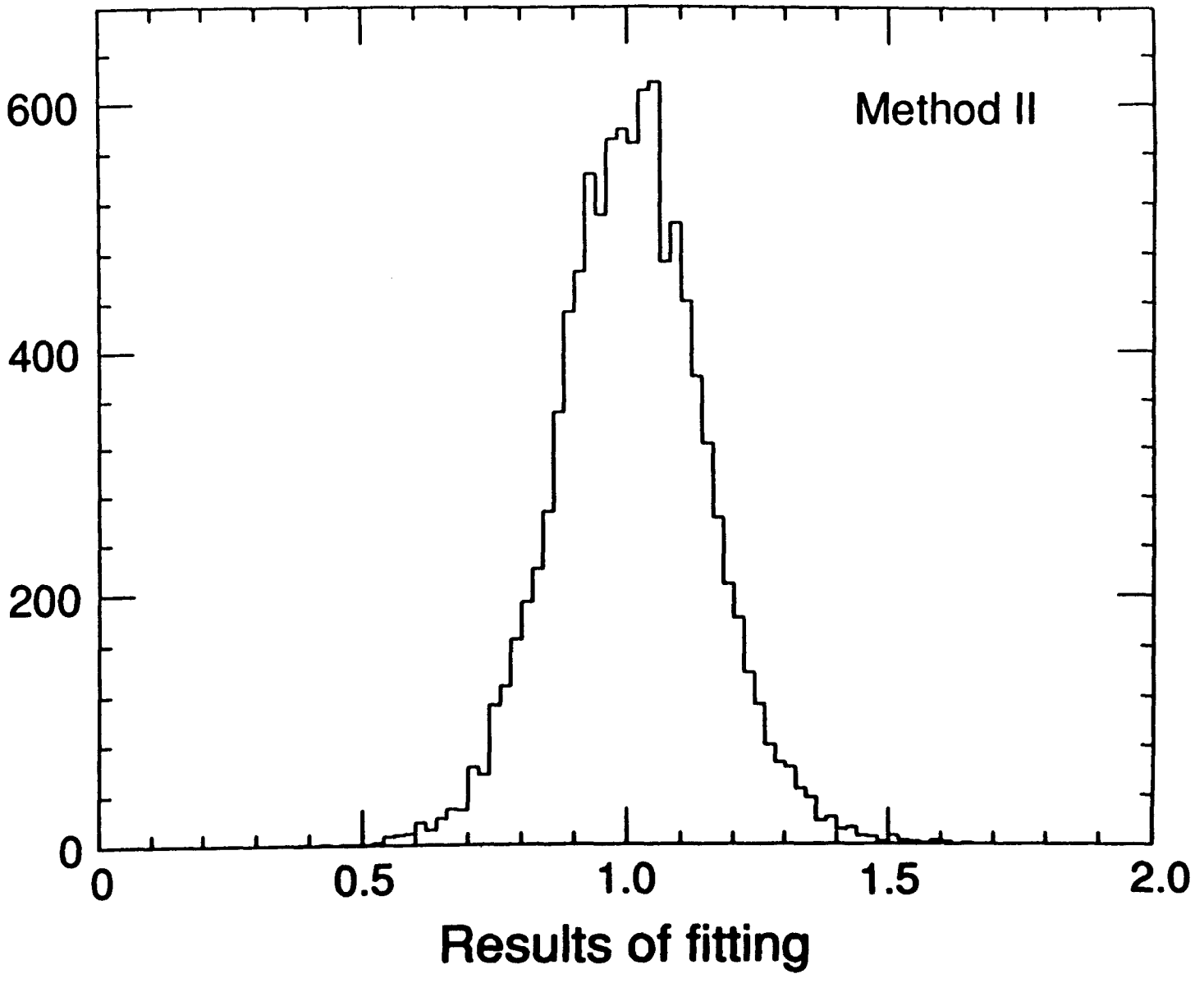


Fig. 5

XBL 898-7708

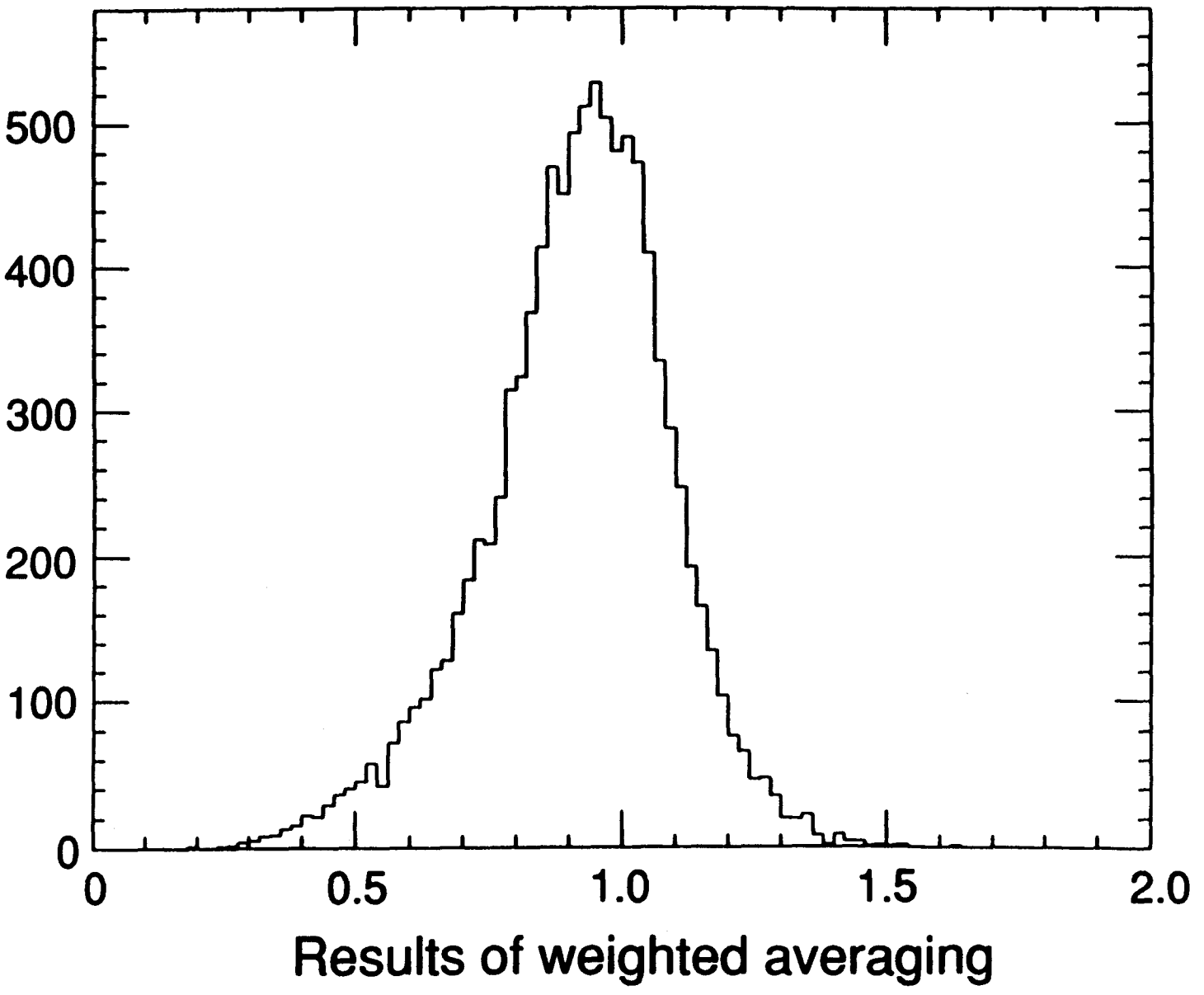


Fig. 6

XBL 898-7707

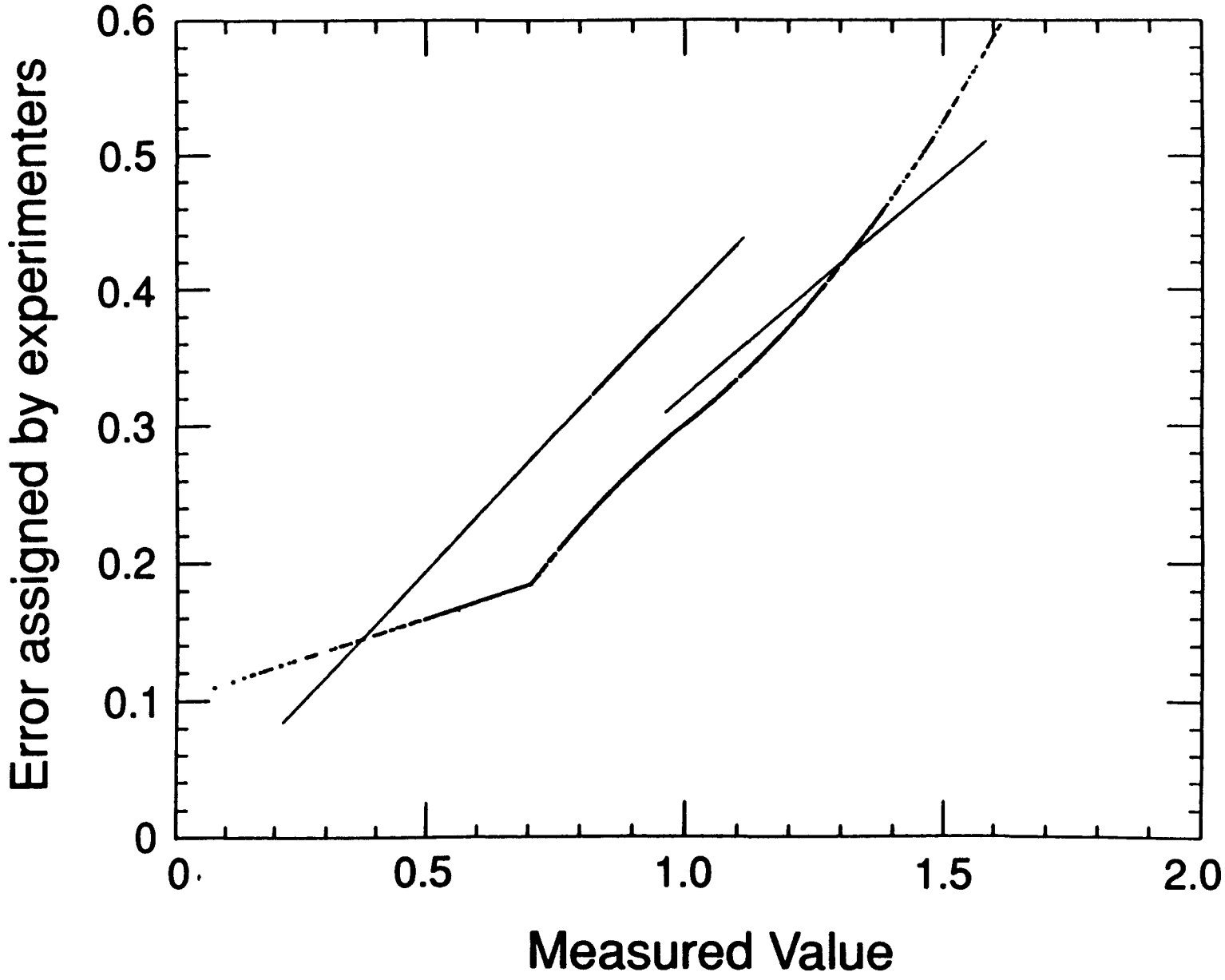


Fig. 7

XBL 898-7713

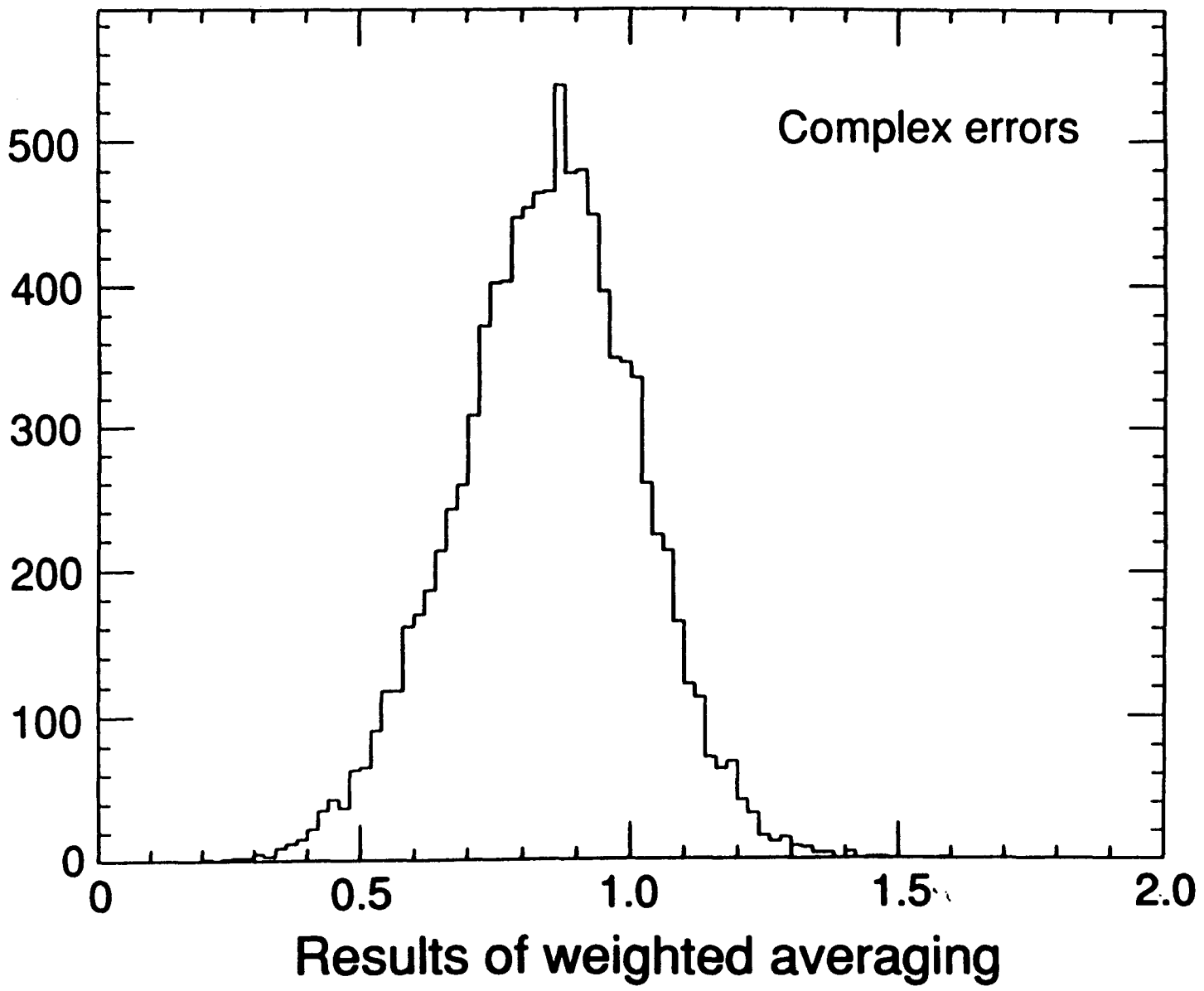


Fig. 8

XBL 898-7706

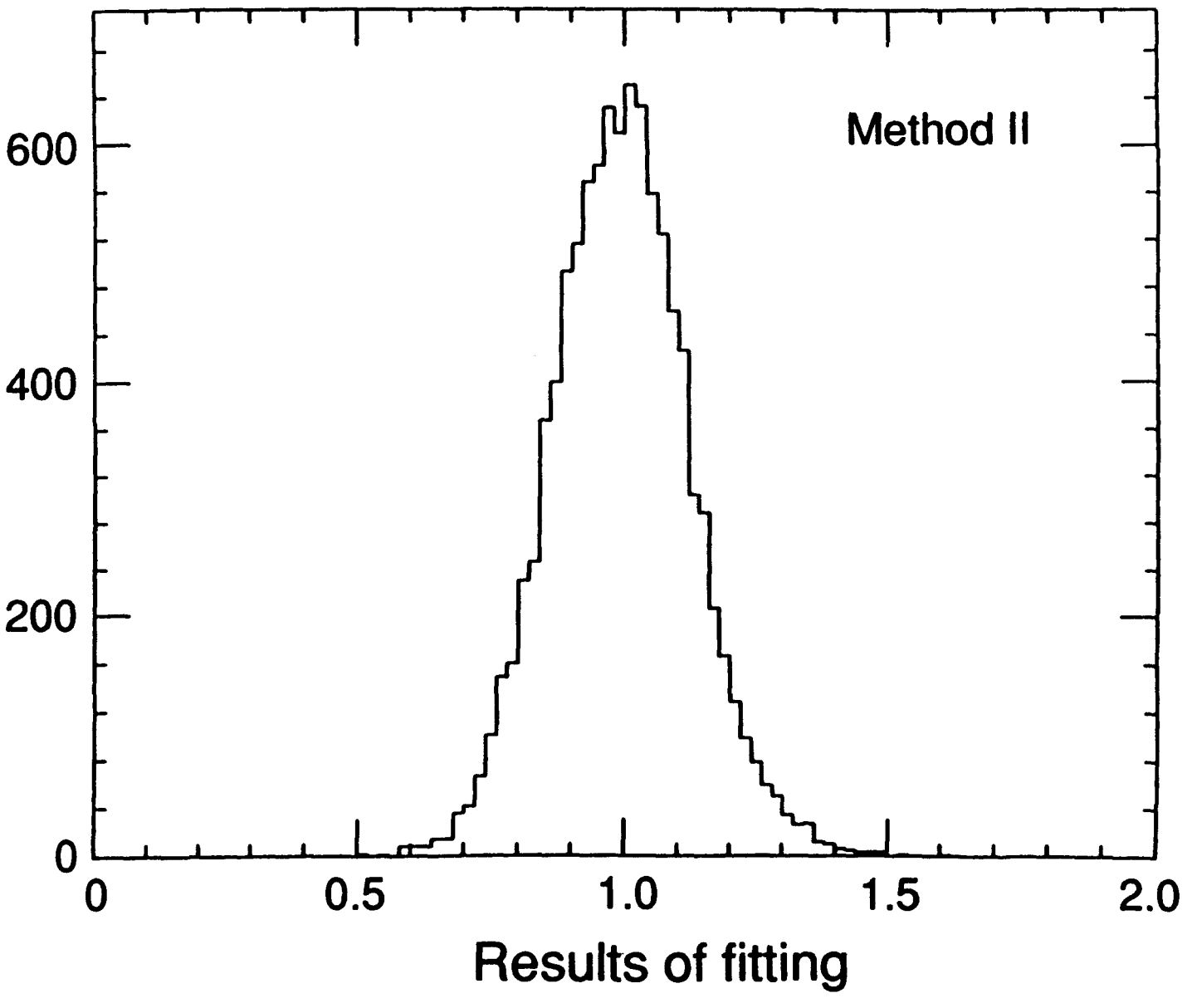


Fig. 9

XBL 898-7714

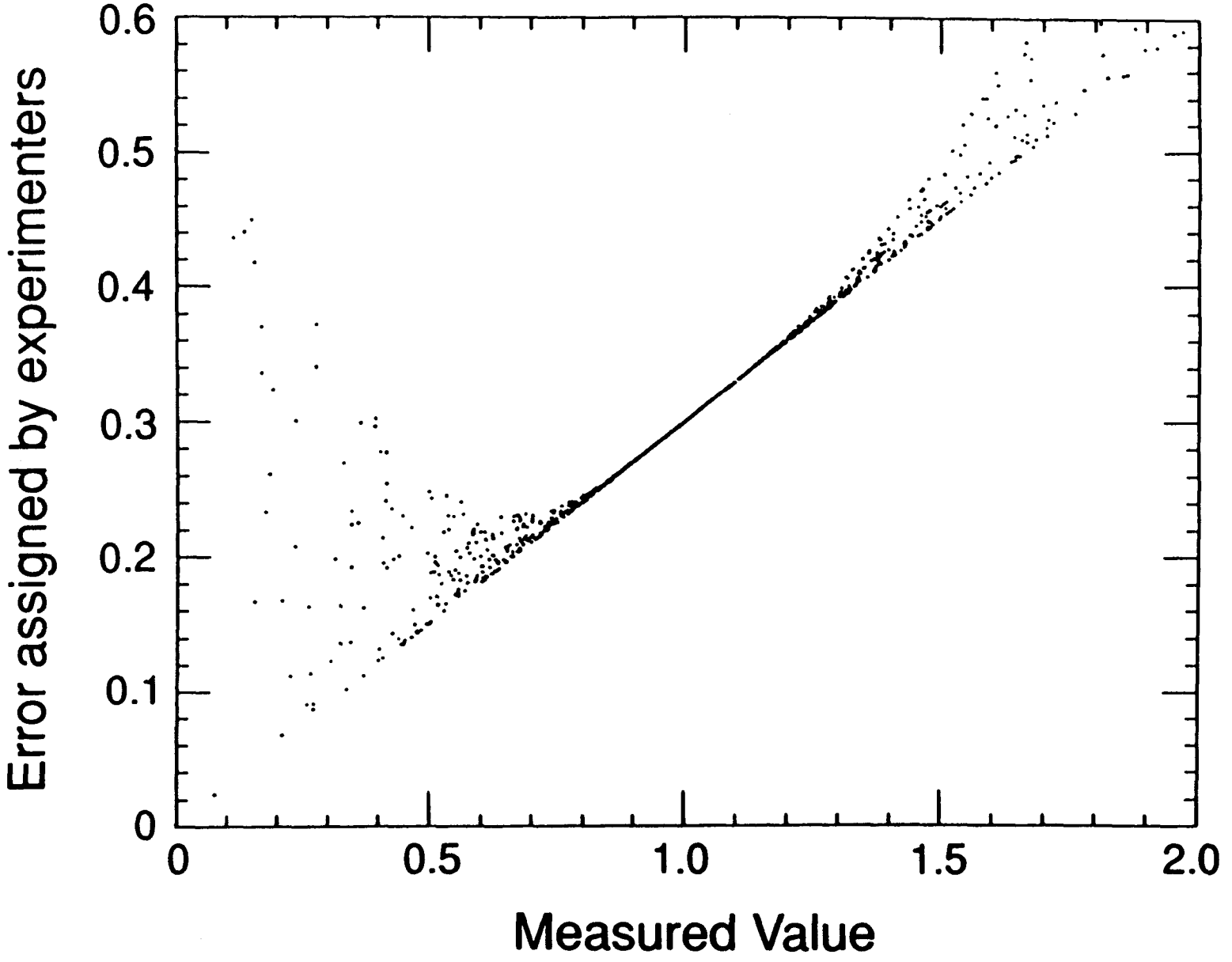


Fig. 10

XBL 898-7717

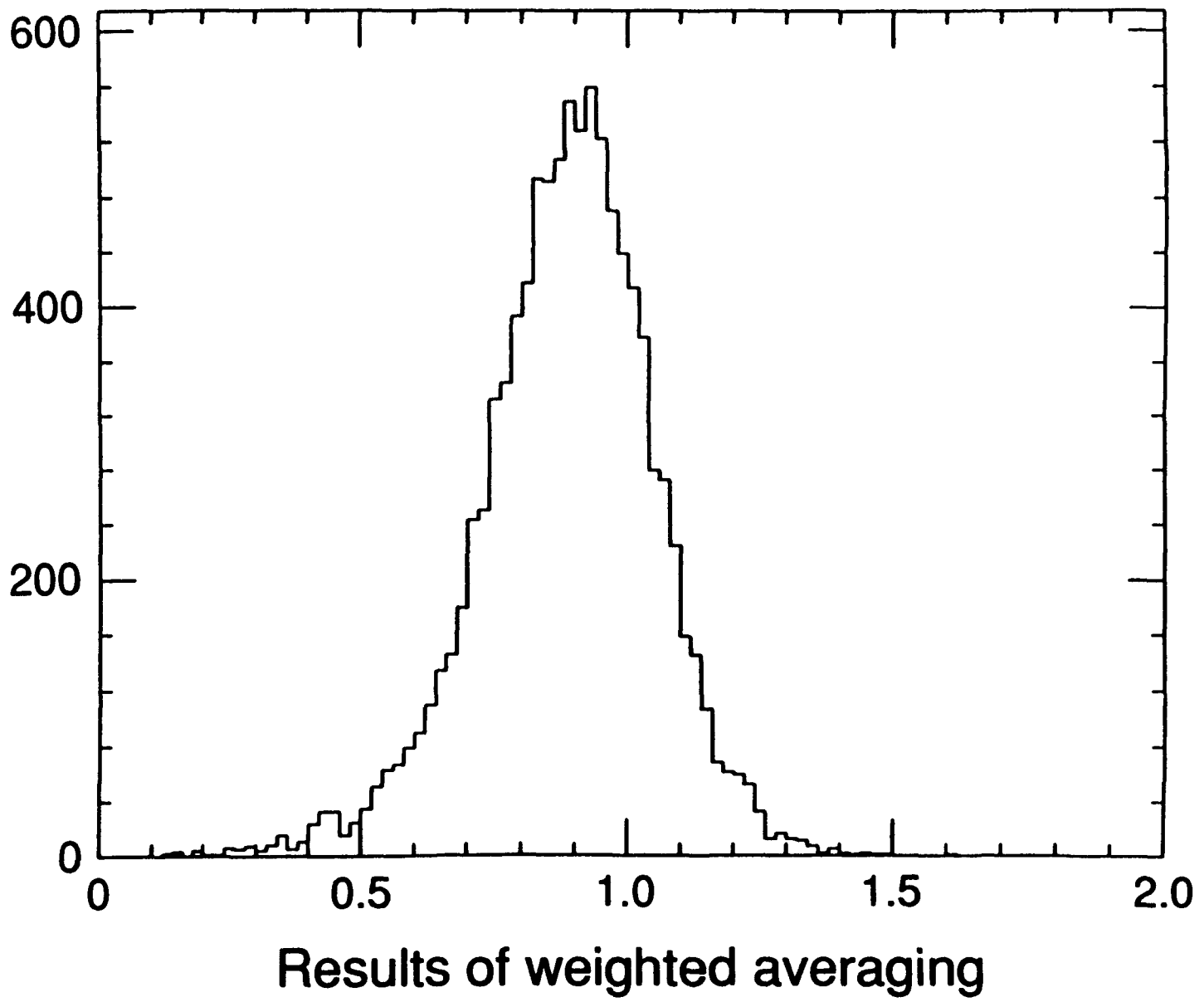


Fig. 11

XBL 898-7716

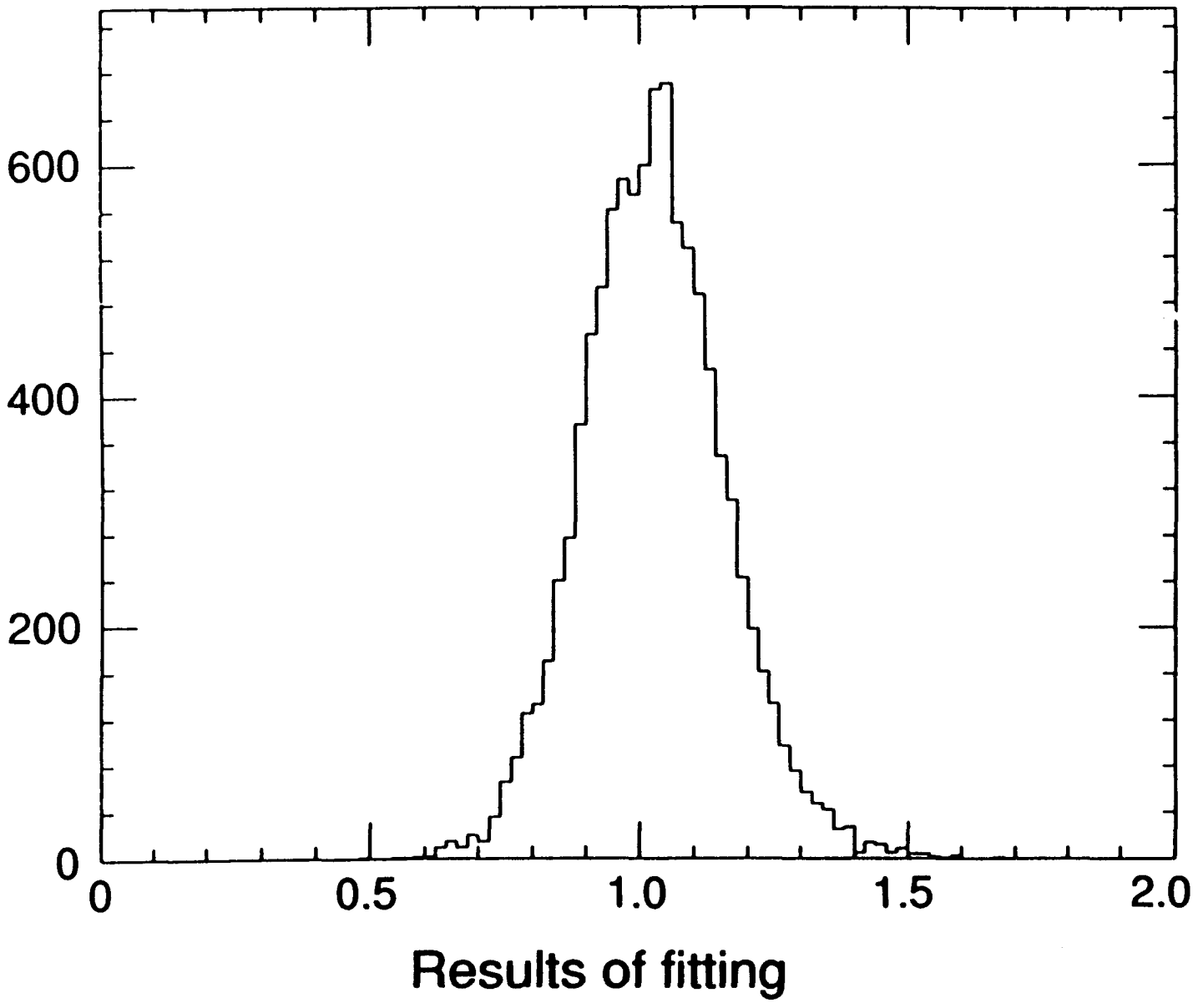


Fig. 12

XBL 898-7720

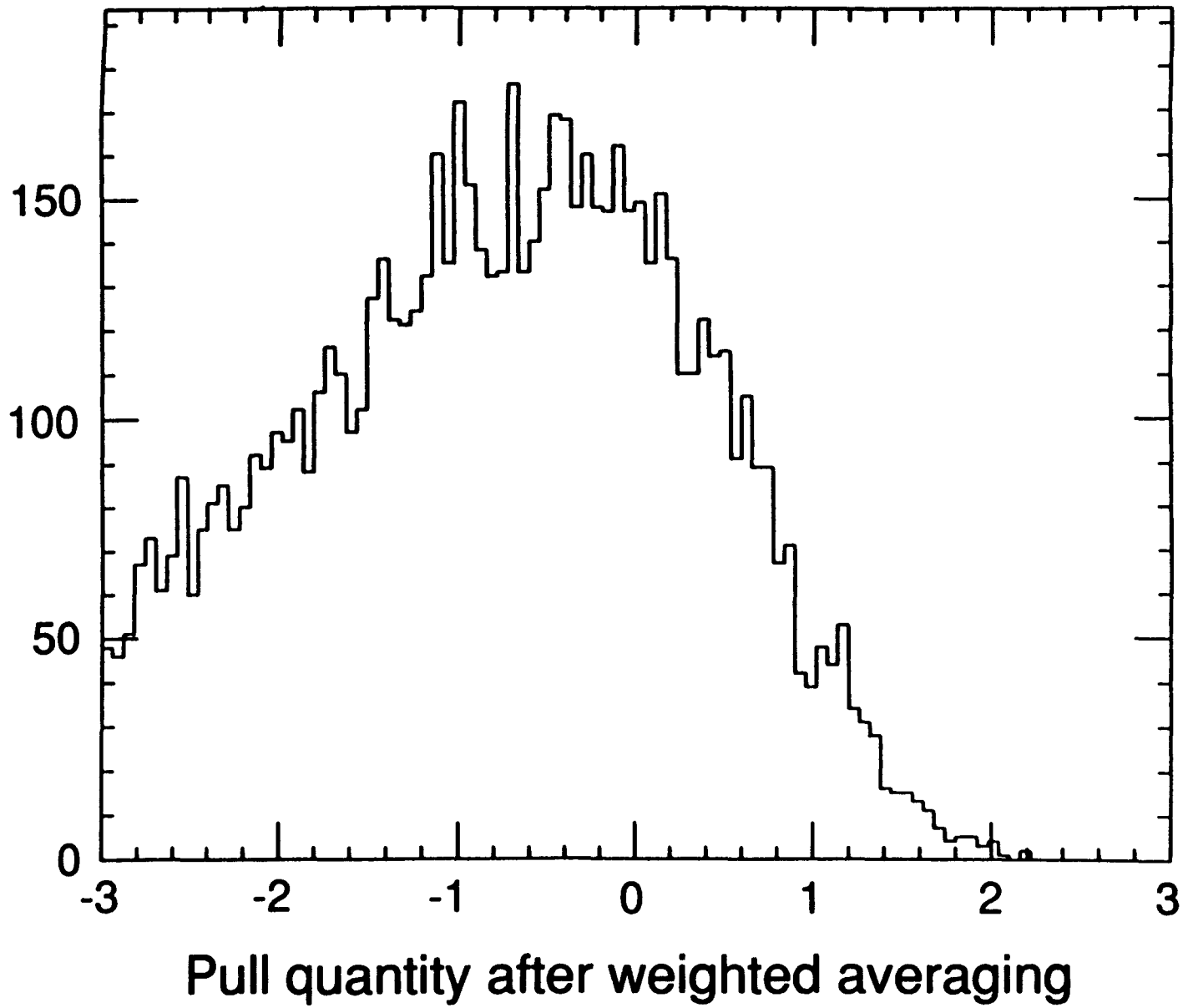


Fig. 13

XBL 898-7719

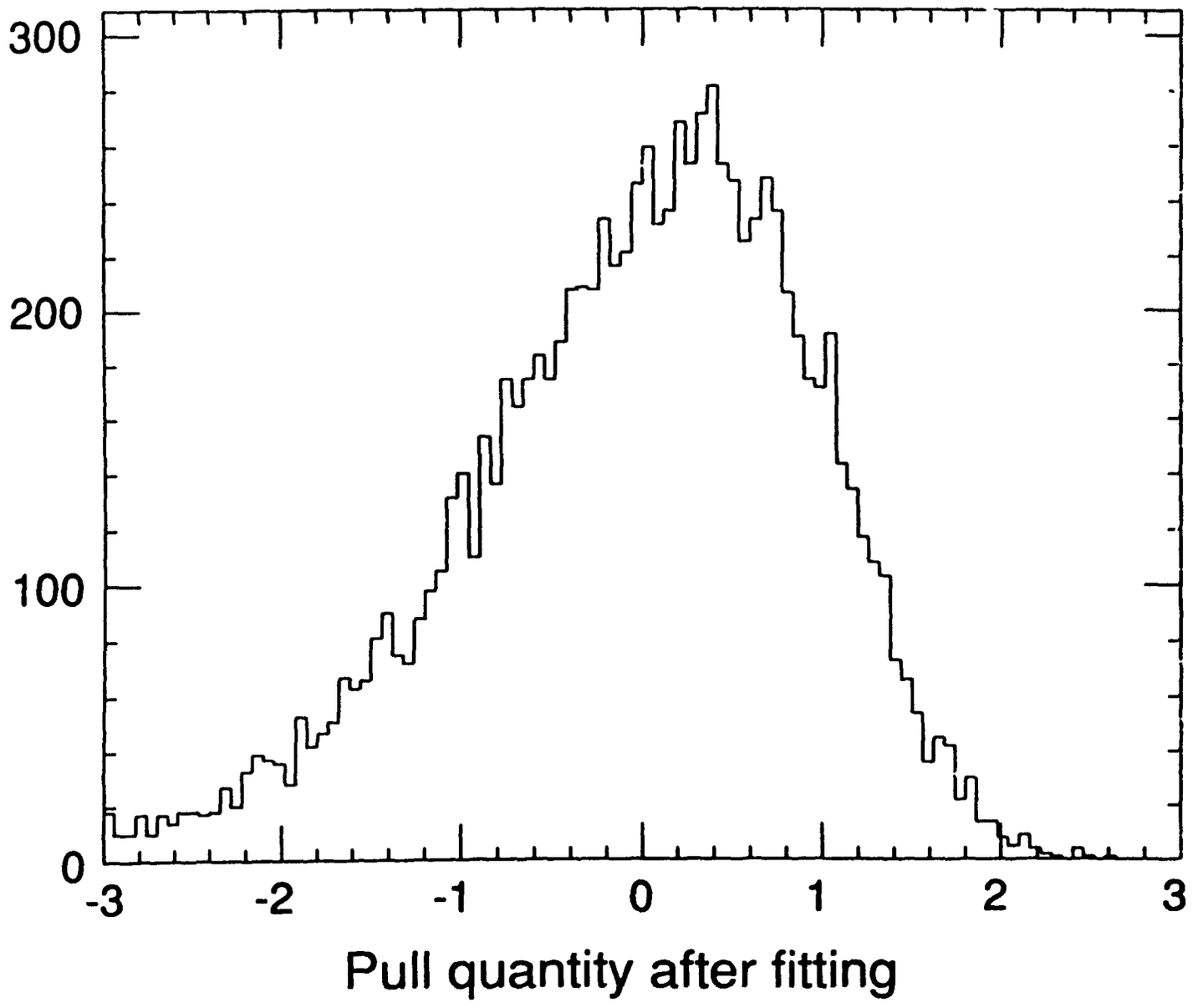


Fig. 14

XBL 898-7715

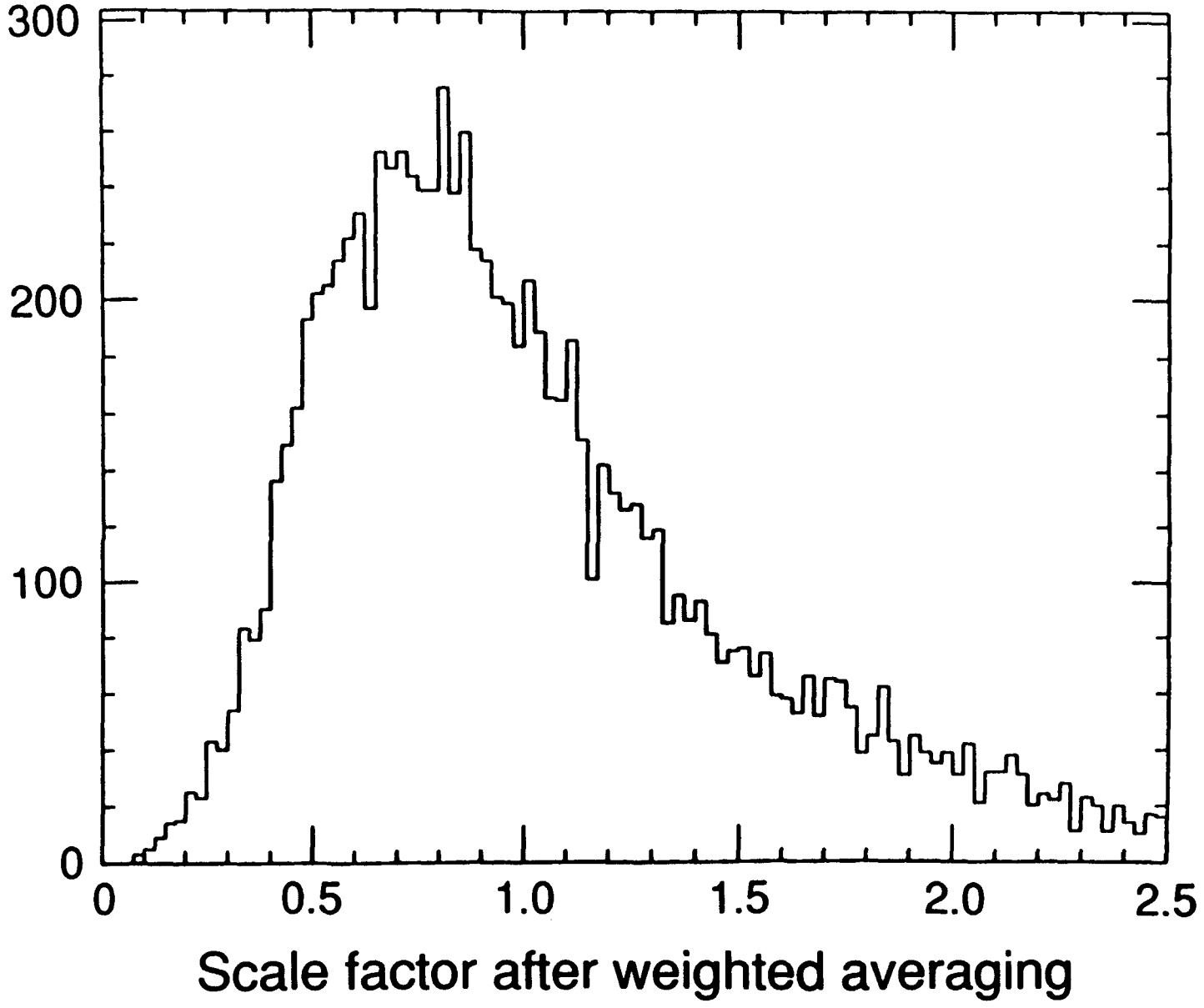


Fig. 15

XBL 898-7721

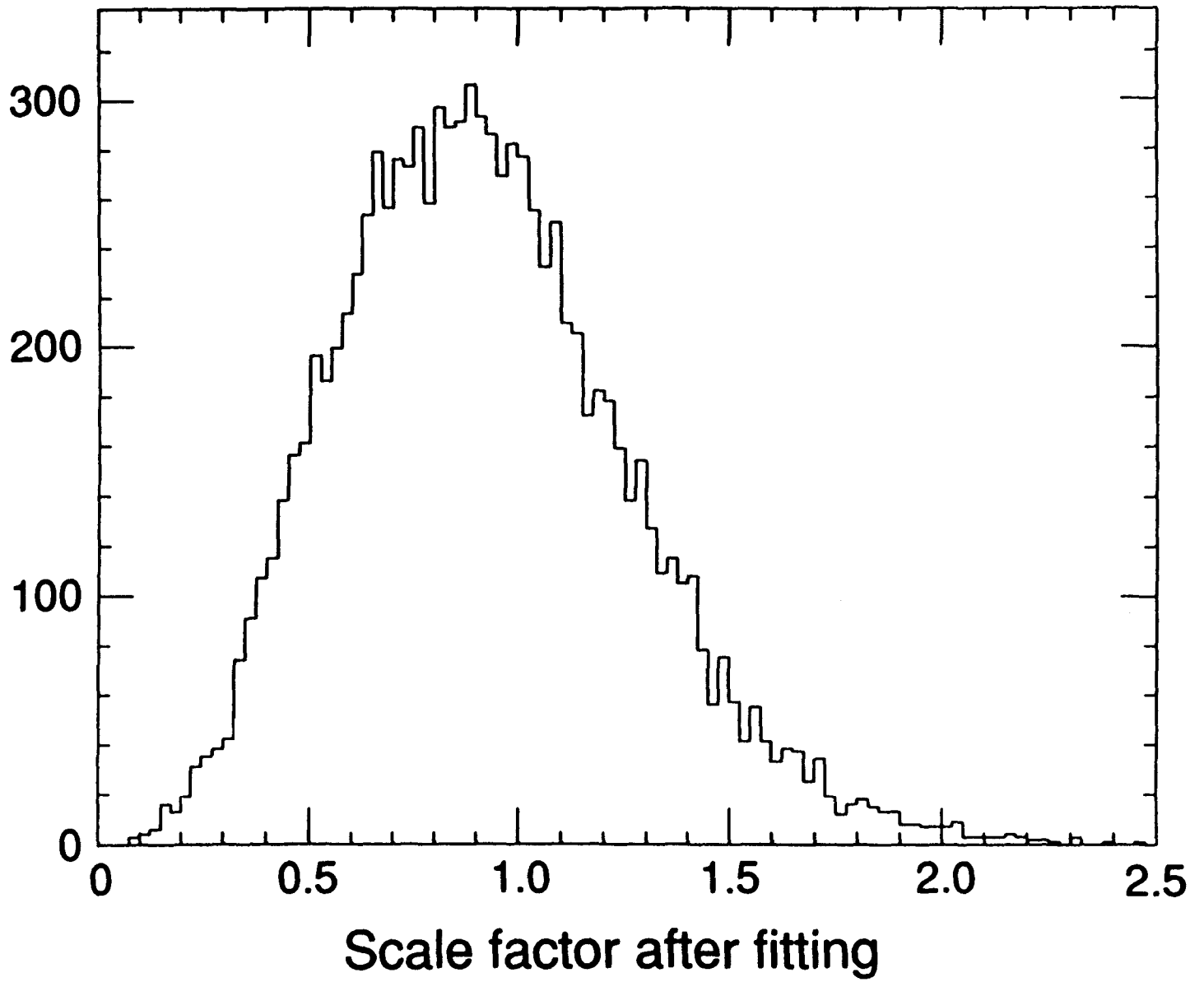


Fig. 16

XBL 898-7718