Chapter 3.1.6

SORTAV (Jun 2008)

Data Merging & Absorption Corrections

R.H Blessing
Hauptman-Woodward Institute
73 High Street
Buffalo
New York 14203 USA
email: mailto:blessing@hwi.buffalo.edu
SOURCE BY ANONYMOUS FTP FROM ftp.hwi.buffalo.edu/pub/Blessing/Drear/

3.1.6 User's Instructions for the Program SORTAV

Program SORTAV is a all-purpose program for treating multiple repeated, symmetry equivalent, and/or azimuth-rotation-equivalent Bragg reflection measurements.

- The program sorts $Y = F^2$ (or $Y = F$) data on *hkl* Miller indices of the equivalent, unique data for the Laue and crystal class point group. The sorted data are ordered so that *h* changes slowest and *l* fastest.
- For data sets measured in subsets on different relative scales (*i.e.*, frames or shells or layers of data, data from different specimen crystals, data at different wavelengths or different incident beam intensities, *etc.*) the program can fit least-squares inter-subset scale factors. [Fox & Holmes (1966), Hamilton *et al.* (1965), Sparks (1970)].
- Given a sufficient redundancy of multiple equivalent or repeated measurements that span a range of specimen irradiation time, the program can derive an empirical correction for Bragg intensity decay due to radiation damage according to

$$Y_{corr}(hkl,i) = Y_{meas}(hkl,i) \times \exp\{\Gamma_{hkl} \times xdose(i)\}$$

  where Γ_{hkl} is an empirically determined correction factor and xdose(i) is proportional to the absorbed dose of radiation up to the time of *i*-th measurement $Y_{meas}(hkl,i)$.
- Given a sufficient redundancy of multiple symmetry-equivalent and/or azimuth-rotation-equivalent measurements, the program can derive an empirical correction for absorption or absorption-like anisotropy by fitting real spherical harmonic functions to the empirical transmission surface as sampled by the multiple equivalent measurements [Blessing, (1995)]
- Replicate and equivalent measurements are averaged according to:

$$Y_{mean} = \frac{\sum_1^n w_i \times Y_i}{\sum_1^n w_i}$$

$$w_i = 1.0 \qquad or \qquad 1/\sigma^2(Y_i)$$

$$esd = \sqrt{\frac{\sum w_i \times \sigma^2(Y_i)}{\sum w_i}} = \sqrt{\frac{\sum \sigma^2(Y_i)}{n}} \quad if \quad w_i = 1.0$$

$$= \sqrt{n/\sum w_i} \quad if \quad w_i = 1/\sigma^2(Y_i)$$

$$rmsd = \sqrt{\frac{n}{(n-1)} \times \frac{\sum w_i \times (Y_i - Y_{mean})^2}{\sum w_i}}$$

$$= \sqrt{\frac{\sum (Y_i - Y_{mean})^2}{(n-1)}} \quad if \quad w_i = 1.0$$

- The $\sigma(\boldsymbol{Y}_i)$ are experimental error estimates as propagated through all the preceding steps of the data reduction process. The program provides schemes for rejecting or down-weighting outliers from the sample median [Blessing (1997), Blessing & Langs (1987)], and the program compiles tables of data merging statistics, which are appropriately adjusted so as to be robust statistics with respect to multiple measurement sample size [Diederichs & Karplus (1997)].
- A bivariate analysis of the variance is performed and the experimental error estimates, esd, are adjusted according to the variation of the ratios rmsd/esd against $\boldsymbol{Y}$-magnitude and $\sin\theta/\lambda$ for the multiple data measurement samples [Blessing (1987), Blessing (1989)].
- Finally, the program estimates the standard uncertainties of the sample means from the root-mean-square measurement uncertainties from the analysis of variance. As described earlier [Blessing, 1987], most error correlation coefficients for Bragg reflection intensity measurements are expected to be positive, and to estimate the standard uncertainty of an *n*-measurement sample mean from the sample estimate of the population standard deviation, one needs an estimate of the mean correlation coefficient, $\langle\rho(\boldsymbol{Y}_i,\boldsymbol{Y}_j)\rangle$. Then, the estimated standard uncertainty of the sample mean is

  $$\sigma(\langle\boldsymbol{Y}\rangle) = \mathrm{sqrt}\{[\langle\sigma(\boldsymbol{Y})^2\rangle/n] \times [1 + (n - 1)] \times \langle\rho(\boldsymbol{Y}_i,\boldsymbol{Y}_j)\rangle\}$$

  It can be shown that, if K_i and K_j represent (possibly implicit) subset scale factors,

  $$\mathrm{cov}(\boldsymbol{Y}_i,\boldsymbol{Y}_j)/(\boldsymbol{Y}_i \times \boldsymbol{Y}_j) = \mathrm{cov}(K_i,K_j)/(K_i \times K_j)$$

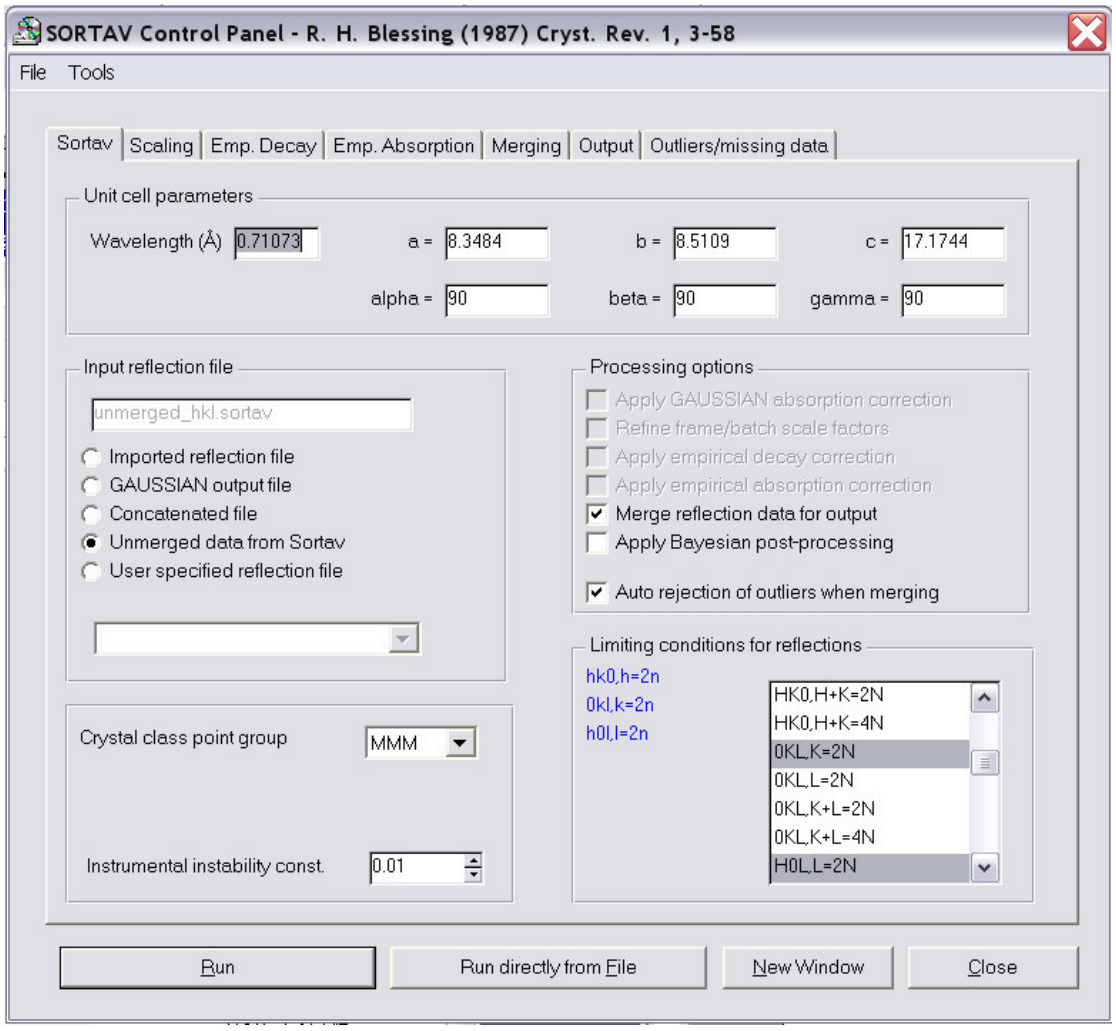
  so that

  $$\rho(\boldsymbol{Y}_i,\boldsymbol{Y}_j)=\rho(K_i,K_j)\times [\sigma(K_i)/K_i] \times [\sigma(K_j)/K_j] \times [\boldsymbol{Y}_i/\sigma(\boldsymbol{Y}_i)] \times [\boldsymbol{Y}_j/\sigma(\boldsymbol{Y}_j)]$$

  Default values, $\langle\rho(K_i,K_j)\rangle = 0.5$ and $\langle\sigma(K_i) \times \sigma(K_j) \;/\; (K_i \times K_j)\rangle = 0.01 \times 0.01$, can be replaced by empirical values from fitting subset scale factors or by user-specified input values.

3.1.6.0 The SORTAVGUI

In the *WinGX* environment, the programs SORTAV & BAYES are run from the SORTAVGUI shown below, which is accessed from the *WinGX* menu item Data-Area Detectors-Sortav. The user needs to select an input reflection file before the Run and Processing option are available (non-greyed). The only reflection file format accepted is the "sortav" free format, where files are called "*_hkl.sortav". In general, processing options which are not available are greyed out (*e.g.* absorption corrections are only available if the input file has direction cosines, batch scaling is only available if there is more than one scale factor in the file *etc*). The GUI is self explanatory, provided the user is aware of the program options and the meaning of input parameters. These are explained in detail in Section 3.1.6.11 of this manual.



3.1.6.1 Concerning least-squares refinement against averaged data

Note that **esd** and **rmsd** are sample-estimates of the standard deviation of the *population*, not the standard deviation of the sample-estimated *mean*. The estimated standard deviation of the mean should decrease with increasing the sample size, *n*, but the estimated standard deviation of the population should remain constant. If *n* equivalent measurements, Y_i, are truly independent measurements, so that their uncertainties, $\sigma(Y_i)$, are uncorrelated, then the standard deviation of the estimated mean is estimated by $esd/\sqrt{n}$ and $rmsd/\sqrt{n}$. It is, however, to be expected that the measurement errors *are* correlated to some extent, although the correlation coefficients are unknown. Error correlations might be due to crystal misorientation, temperature variations, time-dependent scaling errors, inter-subset scaling errors, uncorrected or over-corrected absorption or absorption-like anisotropy, multiple reflection effects, *etc*.

This means that, if $\langle n \rangle$ is the average number of measurements per unique reflection, then the sample *esd* and *rmsd* values will, on average, overestimate the error in the Y_{mean} values by some factor $1 < F < \sqrt{\langle n \rangle}$. The weights for the least-squares structure refinement calculated as $w = 1/esd^2$ or $w = 1/rmsd^2$ will be, on average, underestimated by $1/F^2$; and at the end of the refinement the goodness of fit z = sqrt(sum(w $\times$ delta^2)/(ndat - npar)) should ideally approach $1/F$ rather than unity. At convergence, z > zmin = 1/sqrt(<n>).

<n>	zmin = 1/sqrt(<n>)
1	1
2	0.707
3	0.577
4	0.5
5	0.447
6	0.408
8	0.354
10	0.316
12	0.289
16	0.25
20	0.224

3.1.6.2 Treatment of gross outliers in inter-subset scaling

The least-squares calculation of inter-subset scale factors is nonlinear, and so is done iteratively. In each cycle after the first, gross outliers are omitted if they have

$$\text{abs}[Y(h,i) - Y(h)/K(i)]/\sigma[Y(h,i)] > \text{ZMAX} \times z$$

where

$$z = \text{sqrt}[\chi^2/(\text{ndat} - \text{npar})]$$

and

$$\chi 2 = \text{sum}(h)\text{sum}(i)\ \{[Y(h,i) - Y(h)/K(i)]/\sigma[Y(h,i)]\}^2$$

is the fit residual from the preceding cycle. The coefficient ZMAX can be user supplied or have a default value of ZMAX = 4 to exclude only extreme outliers.

3.1.6.3 Notes on the empirical anisotropic absorption correction

Among the input variables for the empirical absorption correction are the quantities L0MAX and L1MAX, FMU, RADIUS, TMIN and TMAX, and WA :

- L0MAX and L1MAX are, respectively, the even and odd order limits of the real spherical harmonic $Y_{l,m}$ expansion for fitting the absorption anisotropy.
- FMU is the linear absorption coefficient of the crystal.
- RADIUS is an estimated radius for an "equivalent" spherical crystal.
- TMIN and TMAX are, respectively, the minimum and maximum crystal thickness traversed by the incident beam for a tablet- or plate- or blade-shaped specimen crystal.
- WA is a proportionality constant for weighting an absorption anisotropy restraint residual toward $\langle A_{aniso} \rangle = 1$.

3.1.6.3.1 L0MAX and L1MAX

For a crystal bathed in a homogeneous incident beam, the transmission surface is, in principle, centrosymmetric, since reversal of the beam direction gives the same transmission. This implies that the $Y_{l,m}$ fit should be limited to the even order, l = 2n, functions. Odd order, l = 2n + 1, functions can be used if there is also a problem with an inhomogeneous, quasiparallel beam incident from a crystal monochromator, or with a crystal that is too large or not well enough centered to be uniformly illuminated in all orientations. Recommended values for a first run are L0MAX = 8 and L1MAX = 0.

3.1.6.3.2 FMU

If FMU= 0 is supplied, the program will calculate transmission anisotropy factors only. These will range from less than to greater than unity, and typically average approximately unity. They will have no scattering angle θ dependence.

3.1.6.3.3 RADIUS and TMIN

The radius is used to calculate a spherical crystal part of the overall transmission factor, which is defined to be

$$A = A_{sphere} / A_{aniso}$$

where A_{aniso} is the fitted absorption ansiotropy correction factor (i.e., reciprocal transmission anisotropy factor). The spherical crystal part introduces a θ dependence. If RADIUS is supplied as zero but nonzero values are supplied for FMU and TMIN, the program will estimate radius from

$$A_{sphere} = A_{limit} / A_{max}$$

where $A_{aniso} = e^{(-FMU \times TMIN)}$ and A_{max} is the maximum transmission anisotropy factor, i.e., the reciprocal of the minimum absorption anisotropy correction factor A_{aniso}, calculated during the $Y_{l,m}$ fitting. The estimate of the radius from FMU, TMIN, and A_{max} is obtained by interpolation in the table of

$$A(sphere) = A(mu \times r, \theta).$$

from *International Tables for X-ray Crystallography*, Vol. II. If the radius is estimated from FMU, TMIN, and A_{max}, the user should be careful to verify that a reasonable radius is obtained. An unreasonable radius can be obtained if the equivalent data do not thoroughly sample the transmission paths through the

crystal, and the fitted parameters produce an unreasonably large A_{max}. If both RADIUS and TMIN are supplied as zero, the program computes only an absorption anisotropy correction - essentially an anisotropic scaling - and any scattering angle dependence of the absorption correction is neglected.

3.1.6.3.4 WA and restrained least-squares fitting

The empirical absorption anisotropy fit requires a *substantial* redundancy of symmetry equivalent or azimuth-rotation equivalent measurements in order that there be a thorough sampling of the transmission paths through the crystal, and the input reflection data *must be* Y = F^2, rather than Y = *F*. The residual minimized is the sum of a fit residual and a restraint residual for the absorption anisotropy,

$$\chi^2 = \chi(Y) + (w \times \chi^2(A)),$$

where

$$\chi^2(Y) = \text{sum(h) sum(i=1,n) whi} \times (Yhi \times Ahi - \langle Yhi \times Ahi \rangle)^2,$$

$$\chi^2(A) = \text{sum(h) sum(i=1,n) } (Ahi - 1)^2,$$

$$Ahi = 0.5 \times (A(-u0) + A(u1)),$$

$$a(u) = 1 + \text{sum(l=1,lmax) sum(m=-l,+l) A(l,m)} \times Y(l,m)(u),$$

and -u0 and u1 are unit direction vectors, referred to crystal-fixed orthonormal axes, for the reverse incident beam and the diffracted beam, respectively. The restraint residual is intended to restrain the absorption anisotropy correction factors toward an average value of unity, and prevent unreasonable extreme excursions of the fitted transmission surface in regions not sampled by multiple equivalent data. The terms in the fit residual are weighted by

$$whi = 1/\sigma(Yhi)^2,$$

and the restraint residual is weighted by a constant

$$w = wa/(\langle whi \times (Yhi - \langle Yhi \rangle)^2 \rangle / \langle whi \times Yhi^2 \rangle),$$

where the weighting factor WA controls the tightness of the restraint by scaling the restraint residual relative to the fit residual. A value of WA = 1 should lead to normalized mean-square deviations of approximately

$$\langle (Ahi - 1)^2 \rangle / \langle Ahi^2 \rangle = \langle whi \times (Yhi - \langle Yhi \rangle)^2 \rangle / \langle whi \times Yhi^2 \rangle.$$

The user might need to experiment with several different values of L0MAX (and perhaps L1MAX) and WA. The program will automatically decrease L0MAX (and L1MAX) if a singular normal matrix is encountered, but the user should examine the printed output to see if there are many expansion coefficients a(l,m) with values insignificantly different from zero or many large correlation coefficients, indicating the need for a further decrease in L0MAX (or L1MAX). Choice of an appropriate WA value is a matter of the user's judgement of how tightly the restraint <Ahi> = 1 should apply in the case of the particular data set at hand. In summary, prudent use of the empirical absorption subprogram requires intelligent experimentation with trial values of L0MAX (and perhaps L1MAX), WA, and RADIUS.

3.1.6.3.5 Eigenvalue filtering

Solution of the least-squares normal equations is carried out via Jacobi diagonalization and eigenvalue filtering. This technique allows useful sets of A(l,m) parameter values to be determined even if some parameters are not well

determined by the data or some pairs of parameters are so strongly correlated as to be almost linearly dependent. Pseudoparameters p(i) corresponding to eigenvalues with

$$U(i) < umin \times max[u(i)]$$

are assigned values p(i) = 0 before back-transformation from the diagonalization-rotated p(i) parameter hyperspace to the A(l,m) parameter hyperspace [Spackman & Byrom (1997), Watkin (1994)]. The smaller the value assigned to UMIN, the smaller the number of zeroed pseudoparameters. The eigenvalue filtering factor 1.0×10^{-9} roughly corresponds to machine precision for the difference between $\boldsymbol{U_i}$ and the max $\boldsymbol{U_i}$.

3.1.6.3.6 Data selection variables STLMIN, STLMAX, FSQMIN, FSQMAX, AIMIN and AIMAX

User-supplied variables STLMIN, STLMAX, FSQMIN, FSQMAX, AIMIN, and AIMAX are used to select data for the fit of the empirical absorption anisotropy correction. The selected data must obey the conditions :

- $\sin\theta/\lambda \geq$ STLMIN
- $\sin\theta/\lambda \leq$ STLMIN
- $F^2/\sigma(F^2) \geq$ FSQMIN
- $F^2 \leq$ FSQMAX
- $F^2/\text{median}(F^2) \geq$ AIMIN,
- $F^2/\text{median}(F^2) \leq$ AIMAX

where median(F^2) is the median value in each sample of multiple equivalent measurements. STLMAX can be used to exclude from the fitting high-angle data for which absorption effects are relatively small; FSQMIN can be used to exclude data too weak to carry much information about the absorption anisotropy; FSQMAX can be used to exclude data too strong to be free of anisotropic extinction; AIMIN and AIMAX can be used to exclude measurements that are physically unreasonable extreme outliers from their sample median. Default values are:

STLMIN = 0
STLMAX = 9 $Å^{-1}$
FSQMIN = 3
FSQMAX = 10^{10}
AIMIN = 0.5
AIMAX = 1.5

If positive values are supplied for FMU, TMIN, and TMAX, with TMAX > TMIN, the program will calculate :

$$AIMIN = \exp(-FMU \times TMAX)/Amean,$$
$$AIMAX = \exp(-FMU \times TMIN)/Amean,$$

where

$$Amean = 0.5 \times [\exp(-FMU \times TMAX) + \exp(-FMU \times TMIN)].$$

Alternatively, the user can supply pre-calculated or estimated values for AIMIN and AIMAX.

3.1.6.3.7 Crystal orientation information

The arguments -U0 and U1 for the real spherical harmonic expansion functions are orthonormal components of the reverse-incident- and scattered-beam direction unit vectors referred to crystal-fixed Cartesian axes. The program can calculate the Cartesian U-vector components from either vector components or direction cosines referred to either direct space or reciprocal space crystallographic axes, or from Eulerian diffractometer setting angles (2θ, ω, χ, φ) for each reflection measurement.

The program provides three ways to obtain the setting angles:
1. They may be read for each measurement from the input reflection data file.
2. They may be generated from an orientation matrix supplied in the input control data file. In this case, the setting angles generated are those for bisecting, equatorial geometry with $\varphi = \omega = 0$.
3. They may be generated for bisecting, equatorial geometry from an arbitrarily assumed orientation matrix corresponding to a crystal orientation with the crystal a^* axis parallel to the diffractometer x-axis and c^* parallel to z.

Options (2) and (3) can be used as default approximations to deal with data sets for which crystal orientation information for each reflection measurement is not available, but they cannot give correct results for azimuth-rotated measurements made at other than bisecting, equatorial geometry.

3.1.6.3.8 Limits on the applied absorption anisotropy corrections

In order to prevent correction errors due to wild excursions of calculated transmission surface in directions not sampled by multiple measurements equivalent by symmetry or azimuth-rotation, the applied anisotropy corrections are limited to the range of the corrections actually fitted to the multiple equivalent measurement samples, i.e.,

$$A_{min}(fitted) \leq A(applied) \leq A_{max}(fitted).$$

3.1.6.4 Concerning the choice of unit weights or experimental weights for data averaging

For a first run of program SORTAV to identify gross outliers one should use unit weights. Although the relative error, $\sigma(Y_i)/ Y_i$, is generally larger for small Y_i than for large Y_i, the absolute error, $\sigma(Y_i)$, is generally smaller for small Y_i than for large Y_i. Thus, if experimental weights, $w_i = 1/\sigma(Y_i)^2$, are used for averaging, Y_{mean} is biased toward the small Y_i, *i.e.*, toward the measurements with negative errors. For a "good" sample of n multiple equivalent measurements, one expects approximately the same Y_i and $\sigma(Y_i)$ for all n measurements. Thus, approximately constant weights, *i.e.,* unit weights, are appropriate for averaging. Experimental weights are useful when one is merging data from two or more different experiments - different crystals, wavelengths, scan speeds, etc. - with significantly different average levels of random error. However, experimental weights should be used only after the gross outlier measurements have been identified, and rejected from the data set, based on a preliminary run using unit weights. The input file of control data for the program allows for a list of measurements to be rejected.

3.1.6.5 Treatment of outlier measurements in data averaging

Several options are provided for dealing with outliers in samples of multiple equivalent or replicate measurements:

3.1.6.5.1 Optional rejection of abnormally low outliers from the sample maximum

On the presumption that abnormally large errors of measuring Bragg reflection intensities are more likely to be negative than to be positive, the program permits rejection of measurements with

$$Y_i < Y_{max} - 2 \times q \times \sigma(Y_{max}),$$

where Y_{max} is the sample maximum measurement, $\sigma(Y_{max})$ is its estimated uncertainty, and the coefficient q has a user-supplied value. A suitable starting value is 4.0

3.1.6.5.2. Optional rejection of abnormal outliers from the sample median

The program permits rejection of measurements Yi with

$$abs[Yi - median(Yi)] > t,$$

where, for $Y(1) \le Y(2) \le Y(3) .... \le Y(n)$

$$median(Yi) = Y((n+1)/2) \qquad \text{for odd } n = 2m+1$$
$$= 0.5 \times [Y(n/2) + Y((n/2)+1)] \qquad \text{for even } n = 2m$$

and

$$t = max(c1*Ymedian,\ c2*median[\sigma(Yi)],\ c3*1.25*median[abs(Yi - Ymedian)]*sqrt[n/(n - 1)],\ c4*zcrit(n)*max\{median[\sigma(Yi)],\ 1.25*median[abs(Yi - Ymedian)]*sqrt[n/(n - 1)]\})$$

in which the coefficients c1, c2, c3, and c4 have user-supplied values. For a normal distribution, $n(x,\mu,\sigma)$,

$$\mu = \langle x \rangle$$

and

$$\sigma = \langle (x - \mu)^2 \rangle^{1/2} = 1.25 * \langle abs(x - \mu) \rangle.$$

Reasonable trial values for the rejection test coefficients are c1=0.05, c2=0, c3=0, c4=1. The rationale for these values is as follows:

- c1 = 0.05 because if values of $\sigma(Yi)$ are based only or mainly on counting statistics they might seriously underestimate the population standard deviation for the strong reflections. Thus even Yi values within c2 $\times$ 100% of Ymedian might be rejected by the test against $\sigma(Yi)$. For weak reflections, the $\sigma(Yi)$ are essentially determined by counting statistics, but for very strong reflections, the $\sigma(Yi)$ can be essentially independent of counting statistics.
- c2 = c3 = 0 and c4 = 1 to default to zcrit(n) $\times \sigma$.
- zcrit(n) is the value of $z = abs(delta)/\sigma = abs(x - \mu)/\sigma$ corresponding to a normal probability p = 1/(2*n) that z > zcrit.

n	2	3	4	5	10	20	50	100	300	1000
zcrit	1.15	1.38	1.54	1.65	1.96	2.24	2.57	2.81	3.14	3.48

Chauvenet's criterion holds that data with z > zcrit are sufficiently improbable in a sample of n data to be rejected (Young (1969)]. The user can override the Chauvenet criterion by supplying a negative value for c4. Indeed, any or all of the tests can be suppressed by supplying negative values for the test coefficients. For each sample of n equivalent measurements, the test for rejection is performed only once, using the initial estimate Ymean = Ymedian. Repetition of the test after recalculation of Ymean could lead to eventual rejection of all n measurements. For problem cases in which the scatter of the measurements is so great that all n measurements, or all but one measurement, are rejected by the test, the initial estimates Ymean=Ymedian, esd = median[σ(Yi)], and rmsd = 1.25 × median[abs(Yi - Ymedian)] × sqrt[n/(n - 1)] are retained.

3.1.6.5.3. Optional normal probability down-weighting of outliers from the sample median

The program permits outlier down-weighting based on estimated relative normal probabilities. After optional rejection of abnormal outliers from Ymax and/or Ymedian, the median of the remaining sample is taken to be an initial estimate of the sample mean. The larger of either the median experimental error estimate or the median absolute deviation from the median is taken as an estimate of the sample standard deviation. Then, with

$$\mu = \text{median}(Yi)$$

and, depending on the user's choice, either

$$\sigma = \sigma(Yi)$$

or

$$\sigma = \max(\text{median}[\sigma(Yi)], \; 1.25*\text{median}\{\text{abs}[Yi - \text{median}(Yi)]\}*\text{sqrt}[n/(n-1)]),$$

The relative normal probability of each measurement is estimated as

$$wi = \exp\{-0.5*[(Yi - \mu)/\sigma]^2\},$$

and used as a weight for calculating Ymean, esd, and rmsd.

3.1.6.5.4. Optional robust/resistant Tukey weighting

The program permits outlier down-weighting based on so-called robust/resistant Tukey weights. With initial estimates

$$\mu = \text{median}(Yi)$$

and, depending on the user's choice, either

$$\sigma = \sigma(Yi)$$

or

$$\sigma = \max(\text{median}[\sigma(Yi)],$$

1.25*median{abs[Yi – median(Yi)]}*sqrt[n/(n – 1)]),

Robust/resistant Tukey weights for averaging are calculated as

$$wi = [1 - (zi/zmax)^2]^2, \quad \text{if } zi < zmax,$$
$$wi = 0, \quad \text{if } zi \geq zmax,$$

where

$$zi = (Yi - \mu)/\sigma$$

and zmax is a user supplied value. The robust/resistant weights for several different zmax values are compared with relative normal probability weights in the following table.

Z	exp(-0.5 × z^2)	{1 - min[1, $(z/zmax)^2$]}^{2}		
		zmax=4	zmax=6	zmax=8
0	1.0	1.0	1.0	1.0
1	0.607	0.879	0.945	0.969
2	0.135	0.562	0.790	0.879
3	0.011	0.191	0.562	0.739
4	3.3e-4	0.0	0.309	0.562
5	3.7e-6	0.0	0.093	0.371
6	1.5e-8	0.0	0.0	0.191
7	2.3e-11	0.0	0.0	0.055
8	1.3e-14	0.0	0.0	0.0

Robust/resistant weighting, with a default value ZMAX = 6, is the program default.

Experience suggests that:

- Normal probability down-weighting of outliers is too severe, since real experimental error distributions tend to have longer tails than normal distributions
- Data rejection by the Q test against Ymax or the C1, C2, C3, C4 test against Ymedian seems to offer little or no advantage over robust/resistant down-weighting of outliers

3.1.6.6 Output lists of outlier measurements

The program produces up to two output files that list outlier measurements:

1. REJECT.LST
2. OUTLIER.LST
Each of these is an ASCII file sorted on decreasing values of zi = (Yi - Ymean)/esd.

The *REJECT.LST* file lists measurements that were rejected according to option 1 or 2 or assigned zero weight according to option 3 or 4, as described above. The measurements written to the *OUTLIER.LST* file are discordant measurements with **zi > ZLIMIT** from samples with **rmsd/esd > QLIMIT** where qlimit and zlimit have user-supplied values or the default values (QLIMIT = 4, ZLIMIT = 4). As described

below, some or all of the records from the *REJECT.LST* and/or *OUTLIER.LST* files can be included in the input control data file *SORTAV.INPUT* for a subsequent run of the program to exclude outlier measurements from processing as they are read from the input reflection data file.

3.1.6.7 Analysis of variance

The variation of the ratios q = rmsd/esd is analyzed as a function of Y = F_o^2 (or Y = F_o) and *s* = sinθ/λ in two ways:

1. The unique data are classified in intervals of Y and *s*, and a two-way table function of q(Y, *s*) compiled. The table entries are the values <rmsd/esd> averaged within the (Y, *s*) blocks defined by the Y and *s* intervals. As the table is being compiled, the mean value of q, qmean, and the root-mean-square deviation of the q values from their mean, rmsdq, are evaluated.
2. A quadratic surface,

$$q(Y, s) = (Y\ s\ 1)\begin{pmatrix} a11 & a12 & a13 \\ a21 & a22 & a23 \\ a31 & a23 & a33 \end{pmatrix}\begin{pmatrix} Y \\ s \\ 1 \end{pmatrix}$$

$$q(Y, s) = a11 \times Y^2 + a22 \times s^2 + a33 + (2 \times a12 \times Y \times s) + (2 \times a13 \times Y) + (2 \times a\ 23 \times s)\ ,$$

is fitted to the (Yi, *s*i, qi) data by least-squares to minimize

$$\chi^2 = \text{sum}(wi*(qi - q(Yi, si))^2)$$

with

$$wi = ni - 1,$$

Here ni is the number of multiple equivalent measurements of the i-th unique reflection. The goodness-of-fit,

$$z = \text{sqrt}((\chi^2/\text{sum}(wi)) \times nobs/(nobs - npar))\ ,$$

here npar = 6 coefficients (a11, a22, a33, a12 = a21, a13 = a31, and a23 = a32), and an r-factor,

$$r = \text{sqrt}(\chi^2/\text{sum}(wi \times qi^2))\ ,$$

are evaluated. The esd's are then revised according to

$$\text{revised esd} = \max(esd, q(Y, s) \times esd, rmsd)$$

where q(Y, *s*) is calculated from the fitted surface, if **z < rmsdq**, or looked-up in the table (without interpolation), if **z ≥ rmsdq**. The analysis of variance should be performed when data from a fair sampling of the y and s ranges of the data set have been multiply measured.

3.1.6.8 Program limits

- Most arbitary limits are removed in the current, dynamically allocated version of SORTAV. Effectively, the number of reflections which can be handled is limited only by the amount of physical memory available.
- At most 999 separate input data files are allowed.

3.1.6.9 Input reflection data file(s) required

Any of these files may be used for input

- <*name*>*_HKL.SORTAV* special format used by WinGX for SORTAV
- <*name*>*.DATA* binary files from DREAR programs
- *NAME.HKL* ASCII free formats files containing records as described below.

  nref, h, k, l, angles(4), Y, sigma, xtime, tbar, s0, s1
  nref, h, k, l, angles(4), Y, sigma, xtime
  h, k, l, Y, sigma, esd, rmsd, nmeas, tbar, s0, s1
  h, k, l, Y, sigma, iscale, s0, s1
  h, k, l, Y, sigma, esd, rmsd, nmeas
  h, k, l, Y, sigma, iscale, xtime
  h, k, l, Y, sigma, iscale
  h, k, l, Y, sigma
  h, k, l, Y, sigma, iscale, s0, s1, xtime

The record content is recognised automatically, s0, s1 are direction cosines as defined by the IORIENT parameter, xtime if the cumulated X-ray exposure time. These files are opened and read by subroutines which transfer the following formal variables:

variable	**meaning**
II	measurement serial number
IH	
IK	Miller indices
IL	
Y	F^2 or F
SIGY	$\sigma(F^2)$ or $\sigma(F)$
ISCALE	subset scale factor number
A(11)	absorption correction variables
IEND	end-of-file indicator

The array A(11) can contain the diffractometer setting angles, angles(4), in A(1) through A(4) and the absorption-weighted mean path length and incident and diffracted beam direction vectors, TBAR, S0(3), and S1(3), in A(5) through A(11). Default values of A(i) = 0 (i = 1, 11) are supplied. The quantities TBAR, S0, and S1 are calculated by the absorption correction program ABSORB, and can be carried through program SORTAV for possible subsequent use in an analysis of anisotropic extinction, in which case it makes sense to average only repeated measurements, not symmetry equivalent or azimuth-rotation equivalent measurements. The

diffractometer angles are for use in the empirical absorption correction subprogram of program SORTAV. The order in which the diffractometer setting angles are input is important.

diff.type	a(1)	a(2)	a(3)	a(4)
INT.TAB.	two-theta	omega	chi	phi
BUSING-LEVY	two-theta	omega	chi	phi
P3	two-theta	omega	phi	chi
CAD4	theta	phi	omega	kappa

3.1.6.10 Output files

Program SORTAV writes up to nine formatted ASCII output files:

- *DATA.SORTAV* - merged unique reflections
- *DATA_UNAV.SORTAV* - unmerged reflections
- *SORTAV.HKL* – merged unique reflections in SHELX format (optional)
- *XD.HKL* - merged unique reflections in XD format (optional)
- *SORTAV.LST*- line printer output
- *REJECT.LST* - rejected measurements
- *OUTLIER.LST*- statistical outlier measurements
- *MISSING.LST*- reflections not measured
- *SORTAV.CIF* - short summary of merging statistics in CIF format

Each record in the file *DATA.SORTAV* contains the variables:

variable	**meaning**
IH	
IK	Miller indices
IL	
YMEAN	sum(wi × Yi)/sum(wi), where wi = $1/\sigma(Yi)^2$ or wi = 1
SIGMA(YMEAN)	$\sigma(Y) \times \text{sqrt}\{[1+(n-1) \times \langle\rho(i,j)\rangle]/n\}$ where $\sigma(Y)$ = max(esd,rmsd)
ESD	$\text{sqrt}(\text{sum}(wi \times \sigma(Yi)^2)/\text{sum}(wi))$
RMSD	$\text{sqrt}((n/(n-1)) \times \text{sum}(wi \times (Yi - Ymean)^2)/\text{sum}(wi))$
NMEAS	n
TBAR	mean path length of beams in crystal
S0(3)	reverse-incident beam direction vector components
S1(3)	diffracted beam direction vector components

The SORTAV.LST line printer output file is self-explanatory. The *REJECT.LST*, and *OUTLIER.LST* files list

II, IH, IK IL, ISCALE, YI, σ(YI), NMEAS, ESD, (YI - YMEAN)/ESD

Each of the three files is sorted in order of decreasing $|Y_i - Y_{mean}|$/esd, and some or all of the records from these files can be appended to the *SORTAV.INPUT* control data input file to reject outliers from a subsequent run of the program.

- the *REJECT.LST* file lists measurements rejected before data averaging.
- the *OUTLIER.LST* file lists statistical outlier measurements identified as those Y_i with $|Y_i - Y_{mean}|$/esd > zlimit in samples of three or more measurements for which q = rmsd/esd > qlimit.
- the *MISSING.LST* file lists in order of increasing $\sin\theta/\lambda$ the Miller indices of unique reflections that were not measured.

3.1.6.11 Control data file SORTAV.INPUT

In addition to the standard SORTAV input format, the *WinGX* version of SORTAV also reads a special "compact" format, written by the SORTAVGUI. It is a modification of the standard SORTAV input format, which allows easy editing to change the type of job run by SORTAV. A sample file is shown below. Details of all the input parameters are given in the description of the standard format. The "compact" parameter file is less job dependent than the standard input, and all parameters for all possible jobs can be included. The actual jobs to be executed are indicated by the **JOBRUN** keyword, rather than by any parameter values. An asterix "*" immediately before the job name activates this job. Thus in the example below, only scaling and merging are requested, though legal parameters are present for a decay and absorption correction.

Example 1- compact format

```
JOBTIT <your title>
INPFIL <name of input reflection file>
OUTFIL <name of output reflection file>
OUTPUT shelx xd
JOBRUN  *scale  decay  absorb *merge
HKLCON HK0,H=2N
HKLCON 0KL,K=2N
HKLCON H0L,L=2N
CCLASS MMM
CELPAR  8.3484   8.5109  17.1744  90.000  90.000  90.000
ERSCAL 1.0 0.01
YSCALE nscale 1 ptgrp laue serialn 0 start 0 qmin  3.0 zmax  3.0
DCORRN ptgrp laue nmin 4 qmin  3.0 wmin  0.1 pmax  0.50
ABSORB l0max 6 l1max 6 ptgrp laue diff cad4 orient -8
ABSORB matrix ub11 ub12 ub13 ub21 ub22 ub23 ub31 ub32 ub33
ABSORB wavel 0.71073 fmu  0.5 radius 0.25 tmin 0.2 tmax 0.3
ABSORB errmut 0.010 wa   1.0
ABSORB fsqmin  0.300E+01 fsqmax  0.100E+11 stlmin  0.00 stlmax 9.000
ABSORB amin  0.5 amax  1.5 ipath 0 iplot -1
YMERGE ptgrp xtal iw 1 jw 3 zmax   6.0
YMERGE qq 0.0 c1  0.0 c2  0.0 c3  0.0 c4  0.0
OUTVAR rij   0.50 sisj 0.0010 qlim  3.0 zlim  4.0
OUTVAR iprin 0 jprin 0 jpath 0 unav 0
STHLIM stlmin 0.0 stlmax 10.0
HKLREJ <name of rejections file>
```

Special information :
The **OUTPUT** keyword may have the optional subkeywords "shelx" or "xd", which indicate whether, in addition to the standard merged output file "data.sortav", a SHELX format reflection file ("sortav.hkl") or an XD format reflection file ("xd.hkl") should be written. This option is not available in standard format input.
The **OUTFIL** keyword is optional, the output data file defaults to "data.sortav".
As many **HKLCON** keywords as required may be present, each one describing a single condition limiting reflections.
The **HKLREJ** keyword gives the name of the file containing the list of known bad measurements to be rejected on input. Each record in this file has the serial number and Miller indices ii, h, k l in free format

Standard Input Format

Two example files are give here, followed by a detailed description of the parameters below.

Example # 1

```
TITLE
T6_Ins,_NSLS,_Mar._1999,_xtl._#14
INFILE(S)
1
data.xdose
OUTFILE
data.sortav
HKLCOND
1
hkl,-h+k+l=3n
CRYSTAL CLASS
3
UNIT CELL
81.16, 81.16, 33.67, 90, 90, 120
ERROR SCALE
0, 0
YSCALE
0
Laue
0, 0, 0, 0
DECAY
1
Laue
0, 0, 0, 0
ABSORB
4, 3
Laue
H
2
1
 -0.008521  -0.004450 -0.01048
 -0.0001590  0.006883 -0.01244
  0.02162   -0.01769  -0.01006
0.800, 0, 0, 0, 0, 0, 0
0,     0, 0, 0, 0, 0, 0, 3
YMERGE
Laue
0, 0, 0
0, 0, 0, 0, 0, 0
OUTVAR
0, 0
0, 0, 0, 0, 0, 0
STHLIM
0, 0
HKLREJ
```

Example # 2

```
TITLE
Gramicidin_A.CsCl, Xtl_#2.  R-axis II data
INFILE(S)
4
../../../Set1/Subset1/total.sortav
../../../Set1/Subset2/total.sortav
../../../Set2/total.sortav
../../../Set3/total.sortav
OUTFILE
data.merged.sortav
HKLCOND
3
h00,h=2n
0k0,k=2n
00l,l=2n
CRYSTAL CLASS
222
UNIT CELL
31.079, 31.895, 52.110, 90, 90, 90
ERROR SCALE
0.95, 0.05
YSCALE
4
Laue
1, 0, 0, 0
DECAY
0
Laue
0, 0, 0, 0
ABSORB
0, 0
Laue
H
1
0
1.5418, 0, 0, 0, 0, 0, 0
0,      0, 0, 0, 0, 0, 0, 0
YMERGE
Laue
2, 3, 4.0
0, 0, 0, 0, 0
OUTVAR
0, 0
0, 0, 0, 0, 0, 0
STHLIM
0, 0.34
HKLREJ
```

Keywords are indicated in blue, e.g. YMERGE and must be given exactly as they appear here - the parameters are entered on following lines. Note that blank lines may be inserted *before* a keyword, but not after it. ***The user must supply at least records 1 through 6***. If records 7 through 14 are omitted, the program will supply default values. Records are read as free format, so if they are supplied they must contain ***at least dummy zero values*** for each variable.

Rec	Keyword/Data	Description	Format
1	TITLE	Job title	
1.1	title string		(a)
2	INFILE(S)	Input reflection data file name(s)	
2.1	nfile	nfile = number of input file names to be read	(*)
2.2	file1(i), i = 1, nfile	second (and subsequent) record(s) nfile records, one for each input file name	(a)
3	OUTFILE	Output reflection data for sorted, averaged reflection data	
3.1	file2	file name	(a)
4	HKLCOND	Conditions limiting possible reflections	
4.1	ncond	ncond=number of hkl conditions to be read	(*)
4.2	hcond(i), i = 1, ncond	Second (and subsequent) record(s). ncond records, one for each hkl condition, as described below.	(a)
5	CRYSTAL CLASS	Crystal class point group	
5.1	ptgrp	Character string for point group symbol, as described below	(a)
6	UNIT CELL	Lattice parameters	
6.1	cell(6)	a,b,c,alpha,beta,gamma (six values Å,deg)	(*)
7	ERROR SCALE	Scaling constants for experimental error estimates	
7.1	err1, err2	Scaling constants err1, err2 $\quad var(Y_i) = err1*\sigma(Y_i)^2 + (err2*Y_{median})^2$ total variance = statistical variance + instrumental variance (default: err1 = 1.0, err2=0.01. To select err2=0.0, supply err2= -1.0)	(*)
8	YSCALE	Scale factors for subsets of the data set	

8.1	**nscale**	NSCALE = n the number of different subset scale factors = 0 a single scale factor of unity will be assumed for all data; omit records 8.2 through 8.4.	(*)
8.2	**ptgp**	Specifies Laue point group or crystal class point group for equivalent reflections for subset scale factors fitting. PTGP ="LAUE" or "XTAL"	(a4)
8.3	**ktype,kstart,qmin,zmax**	KTYPE = 0 read subset scale factor serial number with each reflection. = 1 set subset scale factor serial number equal to the reflection data file serial number. KSTART = 0 Assume starting values of unity for all the subset scale factors; omit record 8.4. = 1 Read starting values from records 8.4. QMIN = Threshold for excluding weak reflections from the fit of inter-subset scale factors. (Default: QMIN = 3). Exclude measurements for which YI/SIGMA(YI) .LT. QMIN. ZMAX = Threshold for assigning zero weight to extreme outliers in the scale factors fitting. (Default: ZMAX = 3) root-mean-square error of fit from the preceding cycle. Assign zero weight to measurements for which ABS(YI - YMEAN/KI)/SIGMA(YI) .GT. ZMAX*MAX(Z, 1.0), where Z is the standardized root-mean-square error of fit from the preceding cycle.	(*)
8.4	**(i, scalek(i)), i = 1, nscale**	Fourth (and subsequent) record(s) (two values per record). NSCALE records, one for each scale factor. Supply records 8.4 only if KSTART ≠ 0 was supplied on record 8.3. Scale factors are defined such that Y_{output} = (scale factor) $\times Y_{input}$	(*)
9	**DECAY**	Dose dependent radiation damage decay correction variables	
9.1	**idecay**	idecay = o do not do decay correction idecay =1 do decay correction. if idecay = 0 omit records 9.2 and 9.3	*
9.2	**ptgp**	PTGP="LAUE" or "XTAL" (default "LAUE")	(a)
9.3	**nmin, qmin,wmin,pmax**	NMIN = minimum acceptable number of (xdose_i, y_hkl,i) data points for gamma_hkl fitting QMIN = y/sigma(y) statistical significance threshold	*

		WMIN = minimum width of each local xdose_i sample as a fraction of the global width (xdose_max - xdose_min) PMAX = maximum probability of no (x,y) correlation Default values NMIN = 4, QMIN = 3, WMIN = 0.1, PMAX = 0.5	
10	**ABSORB**	Empirical anisotropic absorption correction variables. To by-pass the empirical absorption correction, set L0MAX = L1MAX = 0. and omit records 10.2 through 10.6	
10.1	**l0max, l1max**	L0MAX is even order limit. (L0MAX $\leq$ 8) in the spherical harmonic expansion $Y(l,m)$; $l = 0, lmax$; $m = -l, +l$ L1MAX is odd order limit. (L1MAX $\leq$ 7)	(*)
10.2	**ptgp**	Specifies Laue point group or crystal class point group for equivalent reflections for Ylm fitting. PTGP="LAUE" or "XTAL"	(a)
10.3	**diff**	(one value) DIFF is diffractometer type H Hamilton's Int. Tab. axes BL Busing and Levy's axes P3 Bruker (nee Siemens, nee Nicolet, nee Syntex) P3 or P4 axes CAD4 Enraf-Nonius kappa axes OTHER other	(a)
10.4	**iorient**	(one value) IORIENT = 1, 2, 3, +/-4, +/-5, +/-6, +/-7, or +/-8. IORIENT designates the type of data to specify the crystal orientation for each reflection IORIENT = 1 Generate setting angles for bisecting, equatorial geometry for a default orientation corresponding to orthogonalized reciprocal space crystal axes parallel to the goniostat Cartesian axes at zero setting angles. IORIENT = 2 Generate setting angles for bisecting, equatorial geometry from a given orientation matrix. IORIENT = 3 Read setting angles with each reflection. IORIENT = +/-4 Read components of incident and diffracted beam direction vectors referred to crystallographic direct space axes. IORIENT = +/-5 Read components of direction vectors referred to crystallographic reciprocal space axes. IORIENT = +/-6 Read components of direction vectors referred to crystal-fixed	(*)

		orthonormal, Cartesian axes. IORIENT = +/-7 Read direction cosines of incident and diffracted beam vectors referred to crystallographic direct space axes. IORIENT = +/-8 Read direction cosines referred to crystallographic reciprocal space axes. If IORIENT is negative, the direction of the reverse-incident beam is specified. If, and only	
10.5	**imatrix**	imatrix = 0 do not read orientation matrix imatrix = 1 read orientation matrix as next 3 records	(*)
10.5.1	**ub11 ub12 ub13**	(three values) only if imatrix = 1	(*)
10.5.2	**ub21 ub22 ub23**	(three values) only if imatrix = 1	(*)
10.5.3	**ubb31 u32 ub33**	(three values) only if imatrix = 1	(*)
10.7	**wavlen,fmu, radius, tmin, tmax, errmut, wa**	WAVLEM = wavelength (Å) Cu K_α 1.5418 Mo 0.7107 Ag 0.5609 FMU - linear absorption coefficient (mm^{-1}) RADIUS - estimated radius of "equivalent" spherical crystal (mm) TMIN - estimated minimum crystal thickness (mm) TMAX - estimated maximum crystal thickness (mm) ERRMUT - estimated fractional error in mu*tbar. Typically, ERRMUT = 0.01. (default: ERRMUT = 0.0) WA - relative weighting factor for the <Ahi> = 1 absorption anisotropy restraint residual (default: WA = 1.0. To set WA = 0.0, supply WA = -1.0)	(*)
10.7	**fsqmin,fsqmax, stlmin,stlmax, aimin,aimax,ipath, iplot**	FSQMIN/FSQMAX are minimum/maximum $F^2/\sigma(F^2)$ for measurements to be included in the Ylm fitting. (defaults: FSQMIN = 3, FSQMAX =1^{10}) STLMIN/STLMAX are minimum/maximum $\sin(\theta)/\lambda$ for reflections to be included in the Ylm fitting. (defaults: STLMIN = 0, STLMAX = 9 $Å^{-1}$) AIMIN /AIMAX are minimum/maximum relative transmission factor for reflections to be included in the ylm fitting. (defaults: AIMIN = 0.5, AIMAX = 1.5). To by-pass the AIMAX/AIMIN data selection testing, supply AIMIN = AIMAX = -1.0	(*)

		IPATH = 0/1. If IPATH=1 write to the output reflection file the estimated tbar values (mm) and the beam direction vector components along unit-length crystal axes. IPLOT=1/2/3 For IPLOT=1 plot azimuth-scans for [100], [010], and [001]. For IPLOT=2 also [110], [-110], [101], [-101], [011], and [0-11]. For IPLOT=3 also [111], [-111], [-1-11], and [1-11]. (default: IPLOT = 1) To omit all plots, supply IPLOT= -1.0	
11	YMERGE	Data averaging control variables	
11.1	**ptgp**	Specifies Laue point group or crystal class point group for equivalent reflections to be averaged. PTGP="LAUE" or "XTAL"	(a)
11.2	**iw, jw, zmax**	IW=1 for unit weights, IW=2 for experimental weights JW=1 for unit weights, JW=2 for relative normal probability weights and JW=3 for robust-resistant Tukey weights. ZMAX = maximum permitted value of z Weights for averaging are w = WI $\times$ WJ where : If IW=1, WI= 1.0 If IW=2, WI=1/$\sigma(Y)^2$ If JW=1, WJ=1.0 If JW=2, WJ= exp(-0.5 $\times$ z^2) If JW=3, WJ= $[1 - (z/zmax)^2]^2$ for z < ZMAX or 0.0 for z >= ZMAX where : z = (Y – median(Y))/σ, σ=max(median[σ(Y)], median{|Y-median(Y)|} / 0.6745, (default values: IW=1, JW=3, ZMAX=6.0)	(*)
11.3	**q, c1, c2, c3, c4**	Coefficients for outlier rejection tests If Q > 0, then measurement Y_i is rejected as an abnormal outlier from the sample maximum if $Y_i < Y_{max} - 2\times Q\times\sigma(Y_{max})$, where Y_{max}= max(Y_i). If C1, C2, C3, or C4 > 0, then measurement Y_i is rejected as an abnormal outlier from the sample maximum if $\|\{Y_i - median(Y_i)\}\| > T$, where	(*)

		T = max{C1×median(Y_i), C2×median[σσ(Y_i)], C3×1.25×median{abs[Y_i - median(Y_i)]}×[n/(n-1)]$^{1/2}$, C4×zcrit(n)×max{median[σ(Yi)], 1.25median{abs[Y_i - median(Y_i)]}×[n/(n-1)]$^{1/2}$}} Default values: Q = C1 = C2 = C3 = C4 = 0. These provide no data rejection by either the Q or the C1,C2,C3,C4 outlier rejection tests, and outlier treatment is controlled by the data averaging weighting scheme specified on record 11.2. Reasonable trial values are: Q=4, C1 = 0.05, C2 = 0, C3 = 0, C4 = 1	
12	**OUTVAR**	Output variables	
12.1	**rij, sisj**	Variables for estimating σ(<Y>) from <σ(Y)^2>$^{1/2}$ RIJ = <ρ(K_i K_j)> SISJ = <σ(K_i)×σ(K_j)/(K_i×K_j)> (default values: RIJ = 0.5, SISJ= 0.01*0.01). To select RIJ = SISJ = 0, enter RIJ = SISJ = -1.0	(*)
12.2	**qlimit, zlimit, iprint, jprint, jpath, iunavg**	Output lists control variables QLIMIT = minimum Q = rmsd/esd to define discordant measurement samples. ZLIMIT = minimum Z = abs(Y - Y_{mean})/esd to define statistical outlier measurements in samples with Q > QLIMIT. IPRINT= -1/0/1/2. For -1 do not list any reflections. For 0 list discordant reflection samples with one or more rejected or statistical outlier measurements. For 1 also list special axial reflections h00, 0k0, 00l, hh0, h0h, 0kk, and hhh. For 2 also list special zonal reflections hk0, h0l, 0kl, hkk, hkh, and hhl. JPRINT = n if n > 0, also list every n-th reflection. JPATH = 0/1 do not/do write to the output reflection file averaged tbar values (mm) and beam direction vector components. IUNAVG =0/1 do not/do write an output file "data.sortav.un-averaged" of un-averaged individual reflections with rejected outliers omitted. Default values: QLIMIT = 3 ZLIMIT = 4 IPRINT = 0 JPRINT = 0 JPATH = 0 IUNAVG = 0 JPATH ≠ 0 makes sense only for averaging only repeated measurements, not symmetry equivalent or azimuth-rotation equivalent measurements.	(*)

14	**STHLIM**	Input sin(θ)/λ shell limits	
14.1	**smin1, smax1**	If SMIN1 ≥ 0 and SMAX1 > SMIN1, then only data in the sin(θ)/λ shell for which SMIN1 ≤ sin(θ)/λ ≤ SMAX1 will be processed. Default is no sin(θ)/λ limits.	(*)
15	**HKLREJ**	Measurements to be rejected when read as input	
15.1	**ji, jh, jk, jl**	Serial numbers and Miller indices of known bad measurements to be rejected. Give one record per measurement	(*)

3.1.6.12 Space group conditions limiting possible reflections

The conditions must be entered in the following form, one HKLCOND record in *SORTAV.INPUT* per condition, with no leading or embedded blanks in the records (see sample input file). Note, in particular, that the cyclic permutation order is used for the h and l indices; e.g., for the condition for an n-glide plane perpendicular to the b-axis, enter **H0L,L+H=2N.**

```
HKL,H+K=2N
HKL,K+L=2N
HKL,L+H=2N
HKL,H+K,K+L=2N
HKL,H+K+L=2N
HKL,-H+K+L=3N
HKL,H-K+L=3N
HKL,H-K=3N
HK0,H=2N
HK0,K=2N
HK0,H+K=2N
HK0,H+K=4N
0KL,K=2N
0KL,L=2N
0KL,K+L=2N
0KL,K+L=4N
H0L,L=2N
H0L,H=2N
H0L,L+H=2N
H0L,L+H=4N
HH(-2H)L,L=2N
H(-H)0L,L=2N
HHL,L=2N(R-AXES)
HHL,L=2N
HKH,K=2N
HKK,H=2N
HHL,2H+L=4N
HKH,2H+K=4N
HKK,2K+H=4N
H00,H=2N
H00,H=4N
0K0,K=2N
0K0,K=4N
00L,L=2N
00L,L=4N
000L,L=2N
000L,L=3N
000L,L=6N
```

3.1.6.13 Point group symbols

The symbol for the point group symmetry over which it is desired to average must be entered in the form given in the following Table.

triclinic	*monoclinic*	*orthorhombic*	*tetragonal*	*trigonal*	*hexagonal*	*rhombic*	*cubic*
1	2	222	4	3	6	R3	23
-1	M	MM2	-4	-3	-6	R-3	M3
	2/M	MMM	4/M	312	6/M	R32	432
		2MM	422	321	622	R3M	-43M
		M2M	4MM	31M	6MM	R-3M	M3M
			-42M	3M1	-6M2		
			-4M2	-31M	-62M		
			4/MMM	-3M1	6/MMM		

Note that 42 symbols are tabulated because there are ten cases of alternative settings of axes for seven of the 32 crystallographic point groups:

-42M	-4M2	
3	R3	
-3	R-3	
312	321	R32
31M	3M1	R3M
-31M	-3M1	R-3M
-6M2	-62M	

The symbols give the locations of the symmetry elements with respect to crystal, not reciprocal, lattice axes, but the distinction is important only for the trigonal groups on hexagonal axes, and for the hexagonal groups. For example, for -3M1, with full symbol -3 2/m 1, the crystal *a* and, therefore, *b* axes are 2-fold axes, and the (x,x,0) crystal planes are mirror planes; the reciprocal lattice *a** and *b** axes, which are perpendicular to the crystal *b* and *a* axes, lie in the mirror planes, and the [h,h,0] axis is a 2-fold axis. For the alternative setting -31M, *a* and *b* lie in the mirror planes, and *a** and *b** along the 2-fold axes.

For the non-centrosymmetric point groups, the unique sector of the reciprocal lattice is chosen such that Bijvoet pairs of reflections differ in the sign of their l-index in all point groups except the point group 2, in which the Bijvoet pairs differ in the sign of their k-index. Thus the output file of unique data will contain Bijvoet pairs hkl and hk-l, or hkl and h-kl, but not Friedel pairs of anti-reflections hkl and -h-k-l. Also, the Bijvoet pairs will not, in general, be adjacent in the output file, which is ordered such that l changes fastest and h slowest.

Point group equivalent reflection indices are transformed to unique reflection indices in subroutine EQUIV. Some of the index transformation codes in this subroutine for the higher symmetry point groups have not been thoroughly tested, and users who have a high symmetry case should check that the pertinent code is correct.

3.1.6.14 References

G. C. Fox and K. C. Holmes (1966) *Acta Cryst.* **20**, 886-891.
W. Hamilton, J.S. Rollet, and R.A. Sparks (1965). *Acta Cryst.* **18**, 129-130.
R.A. Sparks (1970). In *Crystallographic Computing*, edited by F.R..Ahmed, pp. 182-184. Copenhagen: Munksgaard publishers, Ltd.]
R.H. Blessing (1995). *Acta Cryst.* **A51**, 33-38.
R. H. Blessing (1997). *J. Appl. Cryst.* **30**, 421-426
R. H. Blessing and D.A. Langs (1987). *J. Appl. Cryst.* **20**, 427-428
K. Diederichs and P.A. Karplus (1997). *Nature Struct. Biol.* **4**, 269-275. erratum. *Nature Struct. Biol.* **4**, 592.
R.H. Blessing (1987). *Crystallogr. Rev.* **1,** 3-58.
R.H. Blessing (1989). *J. Appl. Cryst.* **22,** 396-397.
M.A. Spackman and P.G. Byrom (1997). *Acta Cryst.* **B53**, 553-564
D. Watkin (1994). **A50**, 411-437.
Hugh D. Young (1969*). Statistical treatment of experimental data*, pp. 78, 162. New York: McGraw-Hill Co.

3.1.6.15 User's Instructions for the Program BAYES

3.1.6.16 Synopsis

Given a set of intensity data, $\boldsymbol{Y}$ and $\sigma(\boldsymbol{Y})$, where $\boldsymbol{Y} = 1/\mathrm{Lp} = \boldsymbol{F}_{meas}^{2}$, the program applies Bayes' probability theorem, as described by French and Wilson (1979), to estimate statistical expectation values for $\boldsymbol{F}_{obs}^{2}$, $\sigma(\boldsymbol{F}_{obs}^{2})$, $\boldsymbol{F}_{obs}$, and $\sigma(\boldsymbol{F}_{obs})$.

Bayesian *A Posteriori* observed distribution	=	*A Priori* Wilson distribution	x normal measurement error distribution
measured mean	=	$Fmeas^2$	
std. dev.	=	$\sigma(F_{meas}^2)$	
A priori mean	=	$\langle F_{meas} s^2/\mathrm{epsilon}\rangle$	local average
variance	=	mean	for acentric distribution
	=	2 x mean	for centric distribution
A Posteriori mean	=	F_{obs}^2	
std. dev.	=	$\sigma(F_{obs}^2)$	

The program also produces locally normalized structure factor magnitudes, $\boldsymbol{E}_{obs}$ and $\sigma(\boldsymbol{E}_{obs})$, where

$$E(\mathrm{hkl}) = F(\mathrm{hkl})/\mathrm{sqrt}[\mathrm{epsilon(hkl)} \times \langle F^2/\mathrm{epsilon}\rangle]$$

3.1.6.17 Required files

The program requires a control data file *BAYES.INPUT* and a reflection data file specified in the control data file. The reflection file should contain all, or at least most of, the unique data, and only unique data, with:

- multiple measurements averaged
- symmetry-forbidden, space-group extinguished reflections removed
- all symmetry-allowed reflections, including weak reflections measured as insignificant above background, included.

In the WinGX environment, the program is run as a post-processing option from the SORTAVGUI, and the input file is written automatically, and deleted after each run. The reflection data file is always the merged "data.sortav" file output by SORTAV, which contains information about how many contributers there were for each reflection

Control data file **BAYES.INPUT** (in free format for numerical data)

Card	Parameter(s)	Description	Format
1	TITLE	Job title	(A)
2	ITYPE	Input file type (first 5 words per record) ITYPE = 0 (free) formatted, ASCII ih,ik,il,Fsq,sigFsq 1 (free) formatted, ASCII ih,ik,il,F ,sigF 2 unformatted, binary ih,ik,il,Fsq,sigFsq 3 unformatted, binary ih,ik,il,F ,sigF In WinGX, itype should always be set to zero	(*)
3	FILE1	Input file name (in WinGX it is data.sortav from a SORTAV run)	(A)
4	LATT	Lattice symbol P, A, B, C, F, I, or R	(A)
5	PTGP	Point group symbol. This must correspond to one given in the Table in Section 5.11. For non-centrosymmetric structures, remember to give the (noncentrosymmetric) point group, not the (centrosymmetric) Laue group, even if Friedel or Bijvoet pairs were not measured or were averaged over the Laue group symmetry.	(A)
6	CELL(6)	Lattice parameters. a, b, c, alpha, beta, gamma	(*)
7	NLOCAL	Number of reflections per $\sin\theta/\lambda$ shell for calculating local averages $Fmeas^2/epsilon>$. (Default: NLOCAL = 100, setting NLOCAL=0 lets the program choose the optimal value)	(*)
8	NOBAYES	Set NOBAYES = 1 if the Bayesian replacement is not to be done (default: NOBAYES = 0) If (NOBAYES $\neq$ 0) then • $F^2 = \max(F^2, 0.0)$ • $\sigma(F^2) = \max(\sigma(F^2), 0.0)$ • $F = sqrt(F^2)$ • *If* $F^2 \geq \sigma(F^2)$ $\sigma(F^2) > 0$ then $\sigma(F) = \sigma(F^2/(2*F)$, otherwise $\sigma(F) = 0.5*\text{sqrt}(\sigma(F^2)$ • If $\sigma(F^2) > 0$ and $\sigma(F) > 0$, then IREJECT=0, otherwise IREJECT =1 In principle, the Bayesian replacement should not be done if, in the earlier steps of data processing, the weak reflection data were rejected or modified according to some threshold cutoff.	(*)

The output reflection data files is a SHELX formatted, ASCII file with records

ih, ik, il, F^2, $\sigma(F^2)$

and also optionally an XD formatted reflection file.

In addition, the files "data.hkl.bayes", data.eee.bayes" and "data.ddd.bayes" may also be saved. The latter two contain normalised *E* values *E*(hkl) sorted in order of the *E*-magnitude or *d*-spacing respecively

$$E(hkl) = F(hkl)/\mathrm{sqrt}[\mathrm{epsilon}(hkl)*\langle F^2/\mathrm{epsilon}\rangle]$$

3.1.6.17 References

1. S. French and K. Wilson (1979). *Acta Cryst*. **A34**, 517-525.