$$\sigma_{I_s}^2 = \sum_{i=1}^{m} G\rho_i + (m/n)^2 \sum_{j=1}^{n} G\rho_j \quad (11.2.5.8)$$

$$= G\left[I_s + I_{\rm bg} + (m/n)(m/n)\sum_{j=1}^{n}\rho_j\right], \quad (11.2.5.9)$$

where $I_{\rm bg}$ is the background summed over all peak pixels. We can also write

$$I_{\rm bg} \simeq (m/n)\sum_{j=1}^{n}\rho_j \quad (11.2.5.10)$$

(this is only strictly true if the background region has *mm* symmetry). Then

$$\sigma_{I_s}^2 = G\big[I_s + I_{\rm bg} + (m/n)I_{\rm bg}\big]. \quad (11.2.5.11)$$

This expression shows the importance of the background $(I_{\rm bg})$ in determining the standard deviation of the intensity. For weak reflections, the Bragg intensity (I_s) is often much smaller than the background $(I_{\rm bg})$, and the error in the intensity is determined entirely by the background contribution.

11.2.5.3. *The effect of instrument or detector errors*

Standard-deviation estimates calculated using (11.2.5.11) are generally in quite good agreement with observed differences between the intensities of symmetry-related reflections for weak or medium intensities. This is particularly true if other sources of systematic error are minimized by measuring the *same* reflections five or more times, by doing multiple exposures of the same small oscillation range and then processing the data in space group *P*1. However, even in this latter case, the agreement between strong intensities is significantly worse than that predicted using equation (11.2.5.11). This is consistent with the observation that it is very unusual to obtain merging *R* factors lower than 0.01, even for very strong reflections where Poisson statistics would suggest merging *R* factors should be in the range 0.002–0.003.

An experiment in which a diffraction spot recorded on photographic film was scanned many times on an optical microdensitometer showed that the r.m.s. variation in individual pixel values between the scans was greatest for those pixels immediately surrounding the centre of the spot, where the gradient of the optical density was greatest. One explanation for this observation is that these optical densities will be most sensitive to small errors in positioning the reading head, due to vibration or mechanical defects. A simple model for the instrumental contribution to the standard deviation of the spot intensity is obtained by introducing an additional term for each pixel in the spot peak:

$$\sigma_{\rm ins} = K\frac{\delta\rho}{\delta x}, \quad (11.2.5.12)$$

where $\delta\rho/\delta x$ is the average gradient and K is a proportionality constant. Taking a triangular reflection profile, the gradient and integrated intensity are related by

$$I_s = \frac{1}{12}\left(x^3 + 3x^2 + 5x + 3\right)\frac{\delta\rho}{\delta x}, \quad (11.2.5.13)$$

where x is the half-width of the reflection (in pixels).

Writing

$$A = \frac{1}{12}\left(x^3 + 3x^2 + 5x + 3\right) \quad (11.2.5.14)$$

gives

$$\sigma_{\rm ins} = (K/A)I_s, \quad (11.2.5.15)$$

where the factor A allows for differences in spot size and K is, ideally, a constant for a given instrument.

The total variance in the integrated intensity is then

$$\sigma_{\rm tot}^2 = \sigma_{I_s}^2 + m\sigma_{\rm ins}^2 \quad (11.2.5.16)$$

$$= G\big[I_s + I_{\rm bg} + (m/n)I_{\rm bg}\big] + m(K/A)^2 I_s^2. \quad (11.2.5.17)$$

A value for K can be determined by comparing the goodness-of-fit of the standard profiles to individual reflection profiles (of fully recorded reflections) with that calculated from combined Poisson statistics and the instrument error term. Standard deviations estimated using (11.2.5.17) give much more realistic estimates than those based on (11.2.5.11), even for data collected with charge-coupled-device (CCD) detectors where the physical model for the source of the error is clearly not appropriate.

11.2.6. Integration by profile fitting

Providing the background and peak regions are correctly defined, summation integration provides a method for evaluating integrated intensities that is both robust and free from systematic error. For weak reflections, however, many of the pixels in the peak region will contain very little signal (Bragg intensity) but will contribute significantly to the noise because of the Poissonian variation in the background [as shown by the $I_{\rm bg}$ term in equation (11.2.5.11)]. Profile fitting provides a means of improving the signal-to-noise ratio for this class of reflection (but will provide no improvement for reflections where the background level is negligible).

11.2.6.1. *Forming the standard profiles*

In order to apply profile-fitting methods, the first requirement is to derive a 'standard' profile that accurately represents the true reflection profile. Although analytical functions can be used, it is difficult to define a simple function that will cope adequately with the wide variation in spot shapes that can arise in practice. Most programs therefore rely on an empirical profile derived by summing many different spots. The optimum profile is that which provides the best fit to all the contributing reflections, *i.e.* that which minimizes

$$R_2 = \sum_h w_j(h)\big[K_h P_j - \rho_j(h)_{\rm corr}\big]^2, \quad (11.2.6.1)$$

where P_j is the profile value for the jth pixel, $\rho_j(h)_{\rm corr}$ is the observed background-corrected count at that pixel for reflection h, K_h is a scale factor and $w_j(h)$ is a weight for the jth pixel of reflection h. The summation extends over all reflections contributing to the profile. The weight is given by

$$w_j(h) = 1/\sigma_{hj}^2, \quad (11.2.6.2)$$

and from Poisson statistics σ_{hj}^2 is the expectation value of the counts at pixel j, and is given by

$$\sigma_{hj}^2 = K_h P_j + \left(a_h p_j + b_h q_j + c_h\right). \quad (11.2.6.3)$$

After Rossmann (1979), the summation integration intensity $I_s(h)$ can be used to derive a value for K_h:

$$I_s(h) = K_h \sum_{j=1}^{m} P_j. \quad (11.2.6.4)$$

In equations (11.2.6.3) and (11.2.6.4), as the profile values P_j are not yet determined, a preliminary profile derived, for example, from simple summation of strong reflections used in the detector-parameter refinement can be used, which will give acceptable weights for use in equation (11.2.6.1).

This method of deriving the standard profile is only appropriate for fully recorded reflections. However, in many cases there will be very few or no fully recorded reflections on each image. In such cases the profile is determined by simply adding together the background-corrected pixel counts from all contributing reflections. In the program *MOSFLM* (Leslie, 1992), the profiles are determined using reflections on, typically, ten or more successive images, so that partials will be summed to give the correct fully recorded profile for the majority of the contributing reflections. Tests carried out using standard profiles derived using only fully recorded reflections and equation (11.2.6.1), or using both fully recorded and partially recorded reflections and simple summation, give data of the same quality as judged by the merging statistics.

The reflection profile changes across the face of the detector, due to obliquity of incidence, changes in the projected diffracting volume and geometric factors. In the *MOSFLM* program, this variation is accommodated by determining several standard profiles (typically nine or 25) for different regions of the detector. When evaluating the profile-fitted intensity for a given reflection, a weighted sum of the nearest standard profiles is calculated to provide the best estimate of the true profile at that position on the detector. For the central regions of the detector there will be four contributing profiles, while at the edges there will be between one and three. The weights assigned to each profile vary linearly with the distance from the reflection to the centres of the regions used in determining the standard profiles. An alternative procedure used in *DENZO* (Otwinowski & Minor, 1997) is to evaluate a new profile for each reflection based on spots lying within a pre-specified radius.

11.2.6.2. *Evaluation of the profile-fitted intensity*

Given an appropriate standard profile, the reflection intensity for fully recorded reflections is evaluated by determining the scale factor K and background plane constants a, b, c which minimize

$$R_3 = \sum w_i(KP_i + ap_i + bq_i + c - \rho_i)^2, \tag{11.2.6.5}$$

where the summation is over all valid pixels in the measurement box. As before,

$$w_i = 1/\sigma_i^2 \tag{11.2.6.6}$$

and

$$\begin{aligned}\sigma_i^2 &= \text{expectation value of the counts at pixel } i \\ &= ap_i + bq_i + c + JP_i.\end{aligned} \tag{11.2.6.7}$$

In order to calculate the weights, the background plane constants and summation integration intensity I_s are evaluated as described in Section 11.2.5, at the same time identifying any outliers in the background. The summation integration intensity is used to evaluate the scale factor J in equation (11.2.6.7) using

$$I_s = J\sum_i P_i. \tag{11.2.6.8}$$

In equation (11.2.6.5), the summation is over all valid pixels within the measurement box. This excludes pixels that are overlapped by neighbouring spots (if any) and any outliers identified in the background region.

Minimizing R_3 with respect to K, a, b and c leads to four linear equations from which K, a, b and c can be determined:

$$\begin{pmatrix} \sum wP^2 & \sum wpP & \sum wqP & \sum wP \\ \sum wpP & \sum wp^2 & \sum wpq & \sum wp \\ \sum wqP & \sum wpq & \sum wq^2 & \sum wq \\ \sum wP & \sum wp & \sum wq & \sum w \end{pmatrix} \begin{pmatrix} K \\ a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum wP\rho \\ \sum wp\rho \\ \sum wq\rho \\ \sum w\rho \end{pmatrix}. \tag{11.2.6.9}$$

The profile-fitted intensity I_p is then given by

$$I_p = K\sum_i P_i. \tag{11.2.6.10}$$

The standard deviation in the profile-fitted intensity is given by

$$\sigma_{I_p}^2 = \sigma_K^2\left(\sum_i P_i\right)^2 \tag{11.2.6.11}$$

$$= \left(\sum_i^N w_i\Delta_i^2/(N-4)\right)A_{KK}^{-1}\left(\sum_i P_i\right)^2, \tag{11.2.6.12}$$

where

$$\Delta_i = (KP_i + ap_i + bq_i + c - \rho_i), \tag{11.2.6.13}$$

N is the number of pixels in the summation and A_{KK}^{-1} is the diagonal element for the scale factor K of the inverse normal matrix (used to minimize R_3).

In the case of partially recorded reflections, it is no longer valid to fit the sum of the scaled standard profile and a background plane to all pixels in the measurement box. Partially recorded reflections can have a profile that differs significantly from the standard profile, with the result that the background plane constants take on physically unreasonable values in an attempt to compensate for this difference. Therefore, for partially recorded reflections, the summation in equation (11.2.6.5) is restricted to pixels in the peak region of the measurement box. Minimizing R_3 with respect to the scale factor K then gives

$$I_p = K\sum P_i \tag{11.2.6.14}$$

$$= (\sum w_iP_i\rho_i - a\sum w_iP_ip_i - b\sum w_iP_iq_i - c\sum w_iP_i)\sum P_i/\sum w_iP_i^2, \tag{11.2.6.15}$$

where all summations are over the peak region only.

It is not possible to derive a standard deviation for partially recorded reflections based on the fit of the scaled standard profile (because partially recorded reflections have a different spot profile). For these reflections, the standard deviation can be calculated using equation (11.2.5.17).

11.2.6.3. *Modifications for very close spots*

In order to apply equation (11.2.6.5), it is necessary to exclude all pixels in the measurement box that are overlapped by a neighbouring spot. This applies not only to the pixels of the reflection being integrated, but also to the pixels of all the reflections used to form the standard profile. Consequently, a pixel should be excluded even if it is only overlapped by a neighbouring spot for one of the reflections used in forming the standard profile. When processing data from large unit cells, this can lead to a very high percentage of the background pixels being rejected and therefore a poor determination of the background plane parameters. In these circumstances, the background plane is determined using only background pixels and excluding only those pixels that are overlapped by neighbours for the reflection actually being integrated. The profile-fitted intensity for both fully recorded and partially recorded reflections is then evaluated in the way described for partially recorded reflections in Section 11.2.6.2, with the summation in equation (11.2.6.15) extending only over

peak pixels. The standard deviation in the intensity for partially recorded reflections is derived from equation (11.2.5.17) as before. For fully recorded reflections, the standard deviation has two components: the first is based on the fit of the scaled standard profile to the reflection profile and the second on the contribution from the background:

$$\sigma_I^2 = \sigma_{\rm prof}^2 + \sigma_{\rm bg}^2 \tag{11.2.6.16}$$

$$= \left[\sum_{i=1}^{m} w_i \Delta_i^2 \Big/ (m-1)\right]\left[\left(\sum_{i=1}^{m} P_i\right)^2 \Big/ \sum_{i=1}^{m} w_i P_i^2\right] + (m/n)\sum_{i=1}^{n}(\rho_i - ap_i - bq_i - c)^2, \tag{11.2.6.17}$$

where m and n are the number of pixels in the peak and background, respectively.

11.2.6.4. *Profile fitting very strong reflections*

For very strong reflections, the background level is very small and equation (11.2.6.15) reduces to

$$I_p \simeq \sum w_i P_i \rho_i \sum P_i \Big/ \sum w_i P_i^2, \tag{11.2.6.18}$$

and the weights are given by

$$w_i \simeq 1/JP_i. \tag{11.2.6.19}$$

Substituting for w_i in (11.2.6.18) gives

$$I_p \simeq \sum \rho_i. \tag{11.2.6.20}$$

As pointed out by Z. Otwinowski (personal communication), this shows that for correctly weighted profile fitting, the profile-fitted intensity reduces to the summation integration intensity for very strong intensities.

11.2.6.5. *Profile fitting very weak reflections*

For very weak reflections, all pixels will have very similar counts and therefore all the weights will be the same. For simplicity, consider the case where the profile fit is evaluated only for the peak pixels, then equation (11.2.6.15) reduces to

$$I_P \simeq \sum P_i(\rho_i - ap_i - bq_i - c)\sum P_i \Big/ \sum P_i^2. \tag{11.2.6.21}$$

The second and third summations in this equation depend only on the shape of the standard profile. This shows that the intensity is a weighted sum of the individual background-corrected pixel counts (rather than a simple unweighted sum, as is the case for summation integration). Because the values of P_i are a maximum in the centre of the spot, this will place a higher weight on those pixels where the contribution of the Bragg diffraction is greatest, and a very low weight on the peripheral pixels where the Bragg diffraction is weakest. In this way, profile fitting improves the signal-to-noise ratio without the risk of introducing any systematic error that may result by simply reducing the size of the peak region for weak spots.

11.2.6.6. *Improvement provided by profile fitting weak reflections*

For very weak reflections, where all the weights w_i are approximately the same, the variance in I_p using equation (11.2.6.21) is given by

$$\sigma_{I_p}^2 = \sum {\rm Var}(\rho_i - ap_i - bq_i - c) P_i^2 \left(\sum P_i \Big/ \sum P_i^2\right)^2. \tag{11.2.6.22}$$

Assuming a flat background and very weak intensity, then from Poisson statistics

$$ {\rm Var}(\rho_i - ap_i - bq_i - c) \simeq G\rho_i, \tag{11.2.6.23}$$

and as ρ_i has approximately the same value (ρ) for all pixels,

$$\sigma_{I_p}^2 = G\rho \sum P_i^2 \left(\sum P_i \Big/ \sum P_i^2\right)^2 \tag{11.2.6.24}$$

$$= G\rho \left(\sum P_i\right)^2 \Big/ \sum P_i^2. \tag{11.2.6.25}$$

The variance in the summation integration intensity is simply

$$\sigma_{I_s}^2 = Gm\rho. \tag{11.2.6.26}$$

The ratio of the variances is thus

$$\sigma_{I_s}^2 \Big/ \sigma_{I_p}^2 = m\sum P_i^2 \Big/ \left(\sum P_i\right)^2. \tag{11.2.6.27}$$

For a typical spot profile, the right-hand side (which depends only on the shape of the standard profile) has a value of 2, showing that profile fitting can reduce the standard deviation in the integrated intensity by a factor of $(2)^{1/2}$.

11.2.6.7. *Other benefits of profile fitting*

11.2.6.7.1. *Incompletely resolved spots*

If adjacent spots are not fully resolved, there will be a systematic error in the integrated intensity which will be largest for weak spots that are adjacent to very strong spots. However, the profile-fitted intensity will be affected less than the summation integration intensity, because the peripheral pixels (where the influence of neighbouring spots is greatest) are down-weighted relative to the central pixels (where the neighbours will have least influence).

Further steps can be taken to minimize the errors caused by overlapping spots. Firstly, when forming the standard profiles, reflections are only included if they are significantly stronger than their nearest neighbours. This will minimize the errors in the standard profiles. Secondly, when evaluating the profile-fitted intensity of a particular reflection, pixels can be omitted if they are adjacent to a pixel that is part of a neighbouring spot (rather than having to be part of that spot).

11.2.6.7.2. *Elimination of peak pixel outliers*

In the same way that outliers in the background region can be identified and rejected (see Section 11.2.5.1.1), it is possible in principle to identify outliers in the peak region of fully recorded reflections as those pixels whose deviation from the scaled standard profile is significantly greater than that expected from counting statistics. This approach works well if the feature that gives rise to the outliers affects only a small fraction of the peak pixels and gives rise to large deviations, and this is the case for some zingers or dead pixels, and for diffraction from small ice crystals when collecting data from cryo-cooled samples.

Another source of outliers is the encroachment of a strong neighbouring spot into the peak region, as discussed in Section 11.2.6.7.1. When dealing with peripheral pixels, the outlier test can be applied to both fully recorded and partially recorded reflections, but a high σ cutoff (*e.g.* 10–20) must be used to avoid rejecting pixels that do not fit the profile simply because they correspond to a partially recorded spot.

11.2.6.7.3. *Estimation of overloaded reflections*

Owing to the limited dynamic range of current detectors, it is common for many low-resolution spots to contain saturated pixels. Providing the saturation level of the detector is known, such pixels can simply be excluded from the profile fitting, allowing a reasonable estimate of the true intensity (except when the majority of the pixels are saturated). A knowledge of the strong intensities is essential for structure solution based on molecular replacement

techniques, and so this is a very useful additional feature of profile fitting.

11.2.6.8. *Profile fitting partially recorded reflections*

Greenhough & Suddath (1986) have shown that when profile fitting is applied to partially recorded reflections this leads to a systematic error in the individual intensities, but there is no systematic error in the total summed intensity. Although their analysis is strictly only applicable to the case of unweighted profile fitting, experience has shown that even when using weighted profile fitting there is no evidence of systematic errors in the summed profile-fitted intensities of partially recorded reflections. This is particularly important as many data sets collected from frozen crystals have few, if any, fully recorded reflections.

11.2.6.9. *Systematic errors in profile-fitted intensities*

The fundamental assumption in profile fitting is that the standard profiles accurately reflect the true profile of the reflection being integrated. Errors in the standard profile will result in systematic errors in the profile-fitted intensities. While these errors will often be small compared to the random (Poissonian) error for weak reflections, this is not necessarily the case for strong reflections, as the systematic error is typically a small percentage of the total intensity. Because the standard profiles are derived from the summation of many contributing reflections, small positional errors in spot prediction will lead to a broadening of the standard profile relative to the profile of an individual spot. The same broadening can occur because of the finite sampling interval in the image, which means that a predicted spot position can lie up to half a pixel away from the centre of the measurement box. This error can be minimized by interpolating the pixel values in the image onto a grid which is centred exactly on the predicted position, but the interpolation step itself will inevitably distort the reflection profile. In spite of these difficulties, providing adequate care is taken to determine the crystal and detector parameters accurately (as mentioned in Section 11.2.2), so that the spot positions are predicted to within a small fraction of the overall spot width, there is no suggestion (from merging statistics at least) for significant systematic error, even in the stronger intensities.

Acknowledgements

I would like to thank Dr A. J. Wonacott, Dr P. Brick and Dr P. R. Evans for many stimulating and critical discussions on all aspects of data integration.