Practice of 2D data treatment in SAS

G Pépy
Research Institute for Solid State Physics and Optics, 1525 Budapest, Pf. 49. Hungary
E-mail: gpepy@laposte.net

Abstract. Many Small Angle Scattering experiments are one dimensional. Meanwhile the data of some experiments are 2 dimensional and the experimentalist may find advantages to make a full 2 dimensional data treatment or cannot avoid it. 2 dimensional data treatments present specific difficulties. The first one is an appropriate display of the whole data and of the most important part of them, in a way significant for the scientist. The second difficulty comes from the bad statistics of data which cannot be grouped together. In the following paper we describe various solutions to these problems, developed through a very long practice in LLB (Saclay, France) and SzFKI (Budapest, Hungary).

1. Introduction
The Small Angle Scattering (SAS) experiment data are usually obtained from a 2 dimensional Position Sensitive Detector (PSD); most of the time they depend of the scattering angle only; therefore the experimentalist is happy to gather these data versus the scattering angle (or scattering vector) in a 1 dimensional file and make the data treatment with one of the many appropriate softwares. However many experiments exhibit anisotropy or features not centered on the incoming beam: they cannot be reduced to 1 dimensional data because the model cannot be split in 1 D functions or it may be advantageous to make a full 2D treatment, for instance to have a single prefactor for the x dependent and the y dependent parts of the model.

Many fields provide anisotropic data: a non limitative list would include soft matter submitted to an external field, either shear, or stress, or magnetic field, of course magnetic materials ranging from nanolayers to nuclear industry alloys, and also stressed metallic alloys. A specific domain includes experiments on single crystals, which display non centered feature rather than centered anisotropy.

The domain of 2D data treatment has already been reviewed comparing the then existing softwares [1]. On the contrary this paper will be devoted to the main problems of 2D data treatment, and illustrated by examples treated with the PXY software developed by the author during a very long time [2].

In the next section we shall review how to display and select 2D data in order them to make sense.

In the following section we shall remind what means “fitting” a model and how to avoid some traps.

The problem of bad statistics (low counting rates) will be addressed next.

Finally we shall review a full data treatment case.

2. Display of 2D data
The first idea to display 2D data is to draw them in a 3D frame and to rotate the picture in order to appreciate its shape. This is certainly very attractive to the observer. Meanwhile it is a dead end once
one considers the next steps: how to select the most significant pixels and how to compare a model to the experimental data? Therefore we found that a better way to display 2D data is to make a flat map with various colors according to the intensity. For complex pictures a rainbow palette or a single color palette with shades are not appropriate; palettes with colors changing fast versus the intensity will much better enhance details.

At the beginning of a data treatment one will take care of a small part of the data, because it is easier to appreciate the intensity shape and also because the fit calculation will be faster. Indeed the author never found a good way to directly compare a model to 2D data. As a human being the experimentalist is so much used to a 1D representation to compare data points to a model curve that it seems impossible to escape it. For this purpose the flat color map allows easily to select the most relevant pixels in filters. The pixel intensity is then projected on one axis. The result is finally displayed on a specific drawing. Then a 2D model function can be calculated for each of the selected pixel and the projection process applied in the same way. In this manner it is easy to compare the model to the data. One should note that the projection process alters the function shape. For instance if the filter is a narrow rectangle it will be easy to recognize a Lorentzian or a Gaussian function; the shape will more and more differ from the classic shape as the rectangle is widened. The shape becomes not familiar at all if the rectangle includes the beam catcher hole!

Many 2D data exhibit different properties versus x and y axis. Therefore the most common representation uses rectangular filters (figure 1). They may be horizontal and/or vertical, possibly symmetric versus one axis; there may be several of them.

![Figure 1. A fully oriented side chain liquid crystal polymer in the smectic phase (PAXY spectrometer, LLB, Saclay) [3]. The director is along the horizontal axis. On the right is the flat map of all the data. One observes an anisotropic central scattering due to the contrast between protonated and deuterated main chains and a Bragg peak due to the smectic structure. The intensity of the pixels inside two rectangular filters is projected on the relevant axis and displayed on the left. The continuous line is the fitted model; a flat background + a double centered Lorentzian + a Gaussian*Lorentzian for the Bragg peak.](image)

In many other cases a display with sector shape filters (figure 2) is appropriate, notably to exhibit anisotropy.
However it may happen that the classic display of the projected intensity is not enough to appreciate the quality of the fit. In figure 3, while the model is apparently good, the fit did not converge correctly. A representation of the data – calculated intensity (weighted by the uncertainty) exhibits a misfit of the centre position. Using a free centred double Lorentzian (figure 4) allowed a better fit convergence.

In some cases it may be interesting to fit several files together, notably if only a few parameters differ. In figure 5 only the anisotropy changes versus temperature. One should note that a single prefactor is fitted for both files while it would be impossible to get a single value if 4 distinct
1D files were fitted. This feature of the 2D fitting software was developed to make more reliable the observation of the variation of a few parameters in a group of files.

**Figure 5.** A nematic liquid crystal polymer (D11 spectrometer, ILL, Grenoble) at 2 different temperatures [6]. For each file 2 pairs of centro-symmetric sector filters are shown on the right. On the left the black and red data (1st file) exhibit small anisotropy, while the 2nd file, the lowest blue and highest green data, show a large anisotropy. The model is common to both files. Only the widths are independent. The prefactor is the same.

The polar geometry is valuable in very rare cases as shown on figure 6. While apparently attractive the polar geometry is difficult to use in practice. Displays versus various coordinates may be helpful.

**Figure 6.** Fully oriented biaxial micels (PAXY spectrometer, LLB, Saclay) [7]. The model is the product of a Gaussian versus the radius and 2 symmetric Lorentzian versus the angle. This is a rare case when a polar representation is best. Pixels in the filter are shown from different points of view: radius and angle.
Sometimes the full Q range is obtained measuring at several distances or wave-lengths. For a better understanding the files should be embedded as on figure 7.

**Figure 7.** Empty bubbles in tungsten wires (D11 spectrometer, ILL, Grenoble). The bubbles model is a set of ellipsoids with a Schultz-Zimm size distribution; it was necessary to add a centered power law [8]. The measurement took place at 2 different distances, 6 and 28m in order to increase the Q range. On the right, the map was built with 2 “embedded” files. To characterize the scattering 3 rectangular filters were necessary; the grouped pixels are displayed on the left.

In the case of X-ray pictures obtained with CCD cameras the pixel number is overwhelming. It is even more important to select the relevant parts of the picture than for pictures from gaz detectors (figure 8).

**Figure 8.** Nanochannels in polycarbonate foils (ID01 beam line, ESRF, Grenoble) [9]. A rough sample surface produces Porod scattering singled out in the red filter. The nanochannel model is a cylinder form factor, product of a Bessel function (green oscillating radial part) and a fast decreasing sine (blue longitudinal part).
3. Fitting process

There are several ways to fit a model to data. A very good introduction to data reduction and error analysis may be found in Bevington [10]. In the PXY software we use the mean-square minimization with steepest descent, which is indeed a very common process. For the clarity of this paper we consider useful to remind the main principles.

The distance between the data and the model is evaluated thanks to a “distance” function. Among several properties such a function should be a positive scalar function. Very often the distance function is a $\chi^2$:

$$
\chi^2 = \frac{1}{N-p} \sum_{i=1}^{i=N} \left( \frac{I_i - Y_i \{P\}^2}{\Delta I_i} \right)
$$

where

- $I_i$ intensity in pixel i
- $N$ number of data points
- $p$ number of free parameters
- $\{P\}$ set of parameters
- $Y_i$ calculated intensity in pixel i
- $\Delta I_i$ uncertainty of $I_i$

If the data probability law is a normal law $\chi^2$ decreases down to 1 as the fit converges. This property is very useful to appreciate the quality of a fit. In particular it helps to detect systematic errors due to bad correction of the raw data.

The minimization process runs in cycles. Giving increments to the parameters during one cycle one calculates the $\chi^2$ 1st and 2nd derivatives versus the parameters. This allows to build an approximate set of linear equations. Solving this set of equations provides parameter increments. Assuming that the current position in the parameters space is close to the solution means that the $\chi^2$ is close to the bottom and that the obtained increments will drive to it. Hopefully the $\chi^2$ calculated for the new parameter set is smaller than the previous one. One has merely to cycle the process until the $\chi^2$ does not decrease significantly.

Unfortunately the model functions of interest make the minimization process highly non linear. It is therefore important to start from a not too bad set of parameters. For this purpose it is helpful to select narrow data filters which will be broadened for the final fit. Also it is advisable to let free only a part of the parameters while the most obvious (background, centre position) are kept fixed. Indeed the more free parameters the highest the risk the minimization process falls in a “false” minimum from which it is impossible to escape. When this happens (a hint is a $\chi^2$ far from 1), the only solution is to try better with other sets of start parameters…

Other minimization methods are available. One of the most interesting is the MINUIT package [11] developed at the CERN. It was recently included as a variant in the PXY software. The fitting results are the same as with our previous method. Sometimes MINUIT helps reaching a solution, sometimes our previous method is better (probably thanks to the $\chi^2$ 2nd derivatives calculation).

4. The problem of bad statistics

When dealing with 2D data, the problem of bad statistics occurs very often. Bevington [10] proposes two ways to improve the statistics: repeat the experiment increasing the measurement time or smooth out the data. We shall consider the smoothing in a subsequent section: before applying Bevington’s advices one should consider the SAS raw data correction.

4.1. SAS raw data correction

Classically SAS raw data correction includes substracting a reference data file from the sample file and calibration by a specific incoherent scatterer (water, Plexiglas, vanadium) from which the background has been removed.
For each pixel the corrected scattered intensity is:

$$\left( \frac{d\Sigma}{d\Omega} \right) = F_w \frac{I_s/C_s - I_r/C_r}{I_w/C_w - I_c/C_c}$$  \hspace{1cm} (2)

$$C_s = \frac{(L_{\text{SD}})^2}{A_d d_s M_s T_s}$$  \hspace{1cm} (3)

Where $I$ is the intensity and $C$ a coefficient. The index mean and the factors are:

- $s$ sample
- $r$ reference
- $w$ calibration (water…)
- $c$ background (container…)
- $A$ beam area
- $d$ sample thickness
- $M$ monitor
- $T$ transmission
- $L_{\text{SD}}$ sample to detector distance
- $F_w$ normalization prefactor

Is, Ir, Iw, Ic and Cs, Cr, Cw, Cc are the pixel intensities and coefficients for sample, reference, calibration and background, respectively.

Of course bad statistics make the fit difficult because of the large associated uncertainties (error bars). It is trivial to observe that the uncertainty over the corrected file is the sum of the uncertainties of all the elements of equation (2). We shall assume that the coefficients uncertainty is negligible, including the transmissions. We shall therefore examine the usefulness of the correction files which are few and therefore much less time consuming than the data files acquisition. At this point we should remember Bacon’s remark in his famous book about diffraction [12]: “the uncertainty of the correction files should not be worse than that of the measured files”. He was hinting mostly to the problem of background measurement: this should be measured during a much longer time than the data in order to get the same level of uncertainty. In a modern SANS spectrometer at the end of a neutron guide, the background is very low indeed and experimentalists do not want to lose very much of their precious experimental time just to obtain good statistics… Meanwhile the calibration file usually exhibits small error bars; therefore it makes sense to forget the background correction just to avoid increasing a lot the final uncertainty!

The calibration file itself is extremely important. For gas detectors the efficiency of each cell exhibits a slow drift versus time and it depends mostly of the preamplifier gain for the anode and cathode which define the cell. This last remark leads to a very useful trick. If one assumes that the efficiency of a given cell is merely the product of the gain of the anode preamplifier and of the gain of the cathode preamplifier one can build a calibration matrix in the following way: calculate the average intensity of the calibration file, reject all pixels with intensity too far from the average, calculate again the average intensity for the remaining pixels, then calculate the average intensity of each line $A_l$ and each column $A_c$ and normalize them by the overall average intensity, finally build a new calibration matrix whose element $A_{lc}$ will be the product of $A_l$ and $A_c$. For a square matrix with $n^2$ elements the uncertainty of element $A_{lc}$ will be smaller than that of the initial matrix by a factor $2/n$. Of course this is not fully accurate but the fit will be more precise! In most cases this is a very efficient way to reduce the uncertainty.

Neither the reference file should be discarded at once as it sometimes contains useful information. For instance for protonated materials it allows to determine accurately the background. However for the data treatment one has to question whether it is better to prefer a single fitted background parameter. Assuming the error bars of the reference file are of the same order of magnitude as that of the sample files, replacing it by a single background parameter would divide the uncertainty by 2. Most often this is a worthwhile transaction.
4.2. Smoothing
The true probability law of data obtained with gaz detectors is a Poisson law, which differs quite a lot from a normal law at the smallest counting rates. This takes place at the many pixels close to the background. And it has a strong influence on the fitting process: in equation (1) the data uncertainty $\Delta I_i$ is usually approximated by $\sqrt{I_i}$, which is correct for a normal law but introduces a strong disturbance in case of a Poisson law. For instance the fit may converge towards an acceptable shape of the model but a much too high background parameter. To remedy this feature Bevington [10] advises to smooth the concerned data (it is enough to smooth only the smaller counting pixels).

The least disturbing method is the so-called “parabolic” smoothing where a pixel intensity is replaced by a weighted average over the neighbouring pixels. The weight of the relevant pixel is $1/4$, while the weight for the 4 next pixels on the neighbouring lines and columns is $1/8$ and the weight for the other 4 pixels on the diagonal is $1/16$. This is very efficient to improve the fitting process. A single smoothing pass is enough.

Meanwhile this manipulation has a drawback: it subtly changes the statistics and now the $\chi^2$ decreases to a value lower than 1, which invalidates the convergence criterium and complicates the parameter error bar calculation. Papoular [13] pointed to a solution to this problem, namely replacing not only the intensity but also the associated uncertainty. Let $\Delta_j$ be the mean square deviation (uncertainty) of pixel $j$ and $w_j$ the relevant weight, after smoothing the uncertainty of pixel $i$ should be:

$$\Delta_i = \sqrt{\sum_j (w_j \Delta_j)^2}$$

where the sum runs over the pixels as shown on table 1. This restores the behaviour of the $\chi^2$.

4.3. Wild points
Wild points are pixels for which the intensity lies far away from the expected value. They are suspected of being linked to systematic errors. Authors of the FITTER 1 dimensional data treatment software [14] included a specific procedure. Several such procedures were developed in the literature under the name “robust fitting”. Two of them (Huber, Tukey bisquares [15]) are now also included in the PXY software. In equation 1 is modified mutiplying the estimator $e_i$ (6) in order to decrease the influence of the wild points in the $\chi^2$. Equation 1 now becomes:

$$\chi^2 = \frac{1}{N-p} \sum_{i=1}^{N} (w_i e_i)^2$$

where $N$ is the total number of pixels, $p$ the number of parameters of the model and $e_i = I_i - Y_i / \Delta I_i$ the estimator.

The weights $w_i$ are defined as:

| method    | tuning constant | weight function |
|-----------|----------------|-----------------|
| Huber     | $k = 1.345$    | $w = \begin{cases} 1 & \text{for } |e| \leq k \\ k/|e| & \text{for } |e| > k \end{cases}$ |
| bisquare  | $k = 4.685$    | $w = \begin{cases} \left(1 - \left(\frac{|e|}{k}ight)^2\right)^2 & \text{for } |e| \leq k \\ 0 & \text{for } |e| > k \end{cases}$ |

Meanwhile it seems that if systematic errors are carefully avoided these procedures are of limited interest.
5. Running a 2D data treatment
The experimentalist must be aware that no method will ever improve a bad experiment! The first task will be to read the data and use a representation best suited to the experiment. Do not forget to include a smoothing of small intensity pixels. Then a simple filter or multiple filters will be chosen, narrow enough to select the most significant pixels (in order to further limit the fitting time while getting a first approximation of the parameters). The intensity of the pixels inside the filter is displayed on a $I = f(v)$ diagram, where $v$ is one of the variable $x$, $y$, $\rho$, $\theta$. The model function includes at least a background parameter and one or several functions. It is highly recommended to have a first look at the model shape produced by the input parameters; it should not be too far from the data. Once it is satisfactory one runs the fitting program. Hopefully it will converge. A $\chi^2$ far from 1 is an unmistakable signature of a false minimum, one should start again the fit with another set of parameters. The calculation of the parameter uncertainty depends of the $\chi^2$ being close to 1. Most often the results are good and it is not necessary to repeat the fit with a large filter only to somewhat improve the parameter error bars.

6. Conclusion
We have reviewed many practical aspects for a successful data treatment: the choice of the representation of the data and the filters, the criteria to include (or not) some corrections and smoothing of the raw data, not forgetting a trick to improve the calibration statistics, how to take care of wild points and, last but not least, how to identify a false minimum though the $\chi^2$ observation.

7. Acknowledgements
Special thanks go to R Kahn (Laboratoire Léon Brillouin, Saclay) who provided the author with the first fitting kernel. The author is deeply indebted to two main users of the PXY software: L Noirez (Laboratoire Léon Brillouin, Saclay) and A Len (SzFKI). The PXY software was developed thanks to their many and always interesting requests. In the meantime they showed an extraordinary patience for the lengthy developments and an even more extraordinary patience for the many bugs which they contributed to identify and fix. It is also important to note the thorough discussions in Dubna with A Kuklin, A Islamov and A Soloviev about the MINUIT package and the wild points treatment. The play goes on…

References
[1] Pépy G, ECNS99 proceedings, Data imaging introductory course, 1999, KFKI-1999-04/E report
Pépy G http://www-llb.cea.fr/cours/pepy/dl.htm
[2] Pépy G, J. Appl. Cryst. (2007). 40, s433–s438
[3] Keller P, Carvalho B, Cotton J P, Lambert M, Moussa M and Pépy G 1985 J. Physique Lett. 46 L1065-L1071
[4] Vicentini F, Noirez L, Pépy G and Mauzac M 1995 Europhys. Lett. 32 (8) pp657-662
[5] Noirez L and Lapp A 1997 Phys. Rev. Lett. 78 1 pp70-73
[6] Noirez L H. Mendil, L. Noirez,* and P. Baroni, I. Grillo PRL 96, 077801 (2006)
[7] Kiselev M A, Janich M, Lesieur P, Hoell A, Oberdisse J, Pépy G, Kisselev A M, Gapienko I V, Gutberlet T, Aksenov V I 2002 J. Appl. Phys. A 74 S1239-S1241
[8] Len A submitted to Neutron News
[9] Mendil H, Noirez N, Baroni P and Grillo I 2006 PRL 96, 077801
[10] Bevington P R 1969 Data Reduction and Error Analysis for the Physical Sciences (New Y ork: McGraw-Hill)
[11] James F 1994 Geneva http://wwwasdoc.web.cern.ch/wwwasdoc/minuit/minmain.html
[12] Bacon G 1955 Neutron diffraction (Oxford: Clarendon Press)
[13] Papoular R, LLB (Saclay), private communication
[14] Soloviev A, Stadnik A, Islamov A and Kuklin A 2007 Dubna http://wwwinfo.jinr.ru/programs/jinrlib/fitter/docs/html/index.html
[15] Huber P 1981 Robust statistics (New York: Wiley)
Fox J 2002 An R and S-PLUS Companion to Applied Regression (Thousand Oaks: Sage publications)