STRUCTURAL BIOLOGY
ISSN 2059-7983

Determination of Patterson group symmetry from sparse multi-crystal data sets in the presence of an indexing ambiguity

Richard J. Gildea* and Graeme Winter

Diamond Light Source Ltd, Diamond House, Harwell Science and Innovation Campus, Didcot OX11 0DE, England. *Correspondence e-mail: richard.gildea@diamond.ac.uk

Received 8 January 2018
Accepted 20 February 2018

Edited by R. J. Read, University of Cambridge, England

Keywords: Patterson group symmetry; partial data sets; indexing ambiguity.

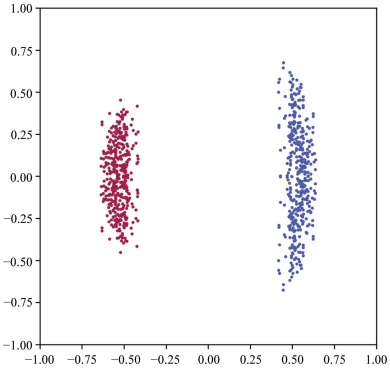
Combining X-ray diffraction data from multiple samples requires determination of the symmetry and resolution of any indexing ambiguity. For the partial data sets typical of *in situ* room-temperature experiments, determination of the correct symmetry is often not straightforward. The potential for indexing ambiguity in polar space groups is also an issue, although methods to resolve this are available if the true symmetry is known. Here, a method is presented to simultaneously resolve the determination of the Patterson symmetry and the indexing ambiguity for partial data sets.

1. Introduction

The recording of an X-ray diffraction data set implies the presence of a crystal lattice with, at the very least, triclinic symmetry. If a relatively complete data set has been recorded from a single crystal, determination of the Patterson symmetry (*i.e.* the symmetry in the diffracted intensities) is relatively straightforward (Evans, 2006); however, this is more challenging for the partial data sets typical of *in situ* experiments, where diffraction data are collected at room temperature. Further complicating matters is the potential for indexing ambiguity in polar space groups, although methods to resolve this are available if the true symmetry is known (Brehm & Diederichs, 2014). Determination of the correct Patterson group is a necessary precondition for the correct scaling of X-ray diffraction intensities. The correct group must be compatible with both the observed crystal lattice and the symmetry in the measured intensities. For substantial data sets the unit cell may be accurately determined and the presence or absence of symmetry operators tested within the single set of observations. For sparse data sets this becomes unreliable within one set, and data sets must be combined before analysis. This, however, depends on correctly matching the data sets to ensure that a consistent setting is used, which in turn requires that the symmetry is known.

The correct crystal symmetry must form a subgroup of the crystal lattice symmetry, although in most cases these are identical. If they are not identical one or more 'twinning operations' exist which map the true symmetry to internally consistent but mutually incompatible cosets within the lattice symmetry group. In contrast to the conventional problem of indexing ambiguity in polar space groups, for sparse data sets accidental ambiguity is more likely, as the uncertainties on unit-cell constants are greater.

Since the symmetry is unknown at the point of integration of the measurements, it may be appropriate to process the



data with a triclinic model and later refine the unit-cell parameters once the symmetry has been determined. This may, however, give rise to up to 24-fold ambiguity if $a \simeq b \simeq c$ and $\alpha \simeq \beta \simeq \gamma \simeq 90^\circ$, in addition to the need to determine the symmetry. Here, we present a method building on that of Brehm & Diederichs (2014) to simultaneously resolve the determination of the Patterson symmetry and the indexing ambiguity for partial data sets. The approach also addresses cases of accidental unit-cell symmetry, *i.e.* lattice pseudo-symmetry such as a monoclinic cell with $\beta \simeq 90^\circ$.

Brehm & Diederichs (2014) introduced a method for resolving the indexing ambiguity from sparse data sets, and a number of implementations of the method, or related approaches, have since been introduced (Gildea *et al.*, 2014; Kabsch, 2014; Ginn *et al.*, 2015; White *et al.*, 2016). Their method is a form of the dimensionality-reduction technique known as multidimensional scaling (MDS). The method uses as input the $n \times n$ matrix of pairwise inter-data-set correlation coefficients, where n is the number of data sets, and outputs a vector, $\mathbf{x}$, of n points in k-dimensional space, where k is generally small (*e.g.* 2 for the case of a twofold indexing ambiguity). In the method presented by Brehm & Diederichs (2014) each data set is used once in its original setting, and thus is represented by a single point in the vector $\mathbf{x}$. They also propose a potential modification of the procedure to include each data set in both its original setting and each of the alternative indexing choices. Here, we present an extension of the methods of Brehm & Diederichs (2014) and Diederichs (2017) to all possible symmetry operations of a given lattice group, allowing simultaneous determination of the Patterson group and resolution of any indexing ambiguity.

2. Methods

2.1. Dimensionality reduction

The maximum possible lattice symmetry compatible with the averaged unit cell is determined using algorithms based on ideas by Le Page (1982) and Lebedev *et al.* (2006), and implemented in *cctbx* (Grosse-Kunstleve *et al.*, 2004; Sauter *et al.*, 2006). Subsequently, a list of all permissible symmetry operations is compiled. The Pearson's correlation coefficient between data sets i and j, after application of the kth and lth symmetry operators, respectively, is defined according to

$$r_{i_k,j_l} = \frac{\sum_h \left[I_{i_k}(h) - \overline{I}_{i_k}\right]\left[I_{j_l}(h) - \overline{I}_{j_l}\right]}{\left\{\sum_h \left[I_{i_k}(h) - \overline{I}_{i_k}\right]^2 \sum_h \left[I_{j_l}(h) - \overline{I}_{j_l}\right]^2\right\}^{1/2}}. \quad (1)$$

The matrix of correlation coefficients is thus a real symmetric matrix, of size $(n \times m)^2$, where n is the number of data sets and m is the number of symmetry operations in the lattice group.

Following Brehm & Diederichs (2014), we represent data sets as coordinates, $\mathbf{x}$, in a multi-dimensional space; however, in this method each data set appears as $n \times m$ coordinates in an m-dimensional space. In the case of pseudo-symmetry, where the true symmetry is $P1$, use of an m-dimensional space is necessary to allow the presence of up to m orthogonal $\mathbf{x}_i$ clusters, where the orthogonality between clusters corresponds to a correlation coefficient r_{i_k,j_l} close to zero.

We then use a modification of algorithm (2) of Brehm & Diederichs (2014) to iteratively minimize the function

$$\Phi = \sum_{i=1}^{n\times m}\sum_{j=1}^{n\times m}\left(r_{i_k,j_l} - \mathbf{x}_i \cdot \mathbf{x}_j\right)^2 \quad (2)$$

using the L-BFGS minimization algorithm (Liu & Nocedal, 1989). As in Brehm & Diederichs (2014), starting coordinates $\mathbf{x}$ are chosen randomly in the range 0–1.

2.2. Principal component analysis

The procedure outlined above in §2.1 is performed in an m-dimensional space, where m is equal to the number of symmetry operators in the lattice group. We anticipate that the points resulting from the minimization above will form a certain number of clusters, given by the ratio of the number of symmetry operators in the lattice group to the number of symmetry operators in the true Patterson group, *i.e.* the number of potential 'twinning' operators. Unless the Patterson group is $P1$, the clusters can be represented in a lower dimensional space that is oriented arbitrarily in the higher dimensional space used for the minimization. Principal component analysis (PCA; Pearson, 1901) may be used to reduce the dimensionality of the resulting clusters of coordinates, which greatly simplifies both the visualization and the further analysis of the clusters. Prior to this analysis, we assume that the true Patterson group, and hence the number of potential twinning operators, are unknown. However, principal component analysis can give an estimate of the relative ratio of the variance of the data that is explained by each principal component, thus giving an indication of the likely number of clusters.

2.3. Symmetry discovery

If the symmetry operator $S_k^{-1}S_l$ is a member of the true Patterson group, then we would expect the coordinates x_{i_k} and x_{j_l} to be part of the same cluster, as the corresponding element of the matrix of correlations, r_{i_k,j_l}, should be close to 1. In contrast, if $S_k^{-1}S_l$ is not a member of the true Patterson group, and thus a potential twinning operator, then we would expect the coordinates x_{i_k} and x_{j_l} to appear in separate clusters, with a correspondingly lower value of r_{i_k,j_l}.

From analysis of a single cluster, it is possible to identify the Patterson group from the combination of all unique symmetry operators $S_k^{-1}S_l$ corresponding to pairs of coordinates x_{i_k} and x_{j_l}. If a potential indexing ambiguity is identified, this can be resolved as follows. If the symmetry operator S_k that corresponds to the coordinate x_{i_k} belongs to the Patterson group determined above, then data set i does not need reindexing. If, however, the symmetry operator S_k does not belong to the Patterson group, then S_k is a twinning operator that can be used to reindex data set i. Analysis of any further clusters should yield identical results.

The reindexing operator determined using the above procedure will be one from a coset of equivalent reindexing operators. This can be mapped to a unique coset representative using left coset decomposition of the lattice group with respect to the proposed Patterson group (Flack, 1987).

3. Results

3.1. Example 1: simulated microfocus data

Diffraction patterns for 100 partial data sets were generated by James Holton (Holton, 2015) from the PDB model of titin (PDB entry 1g1c; Mayans *et al.*, 2001) as an explicit challenge to the community of macromolecular crystallography software developers. The space group of the generated data sets is $P2_12_12_1$, as in the published structure; however, the unit cell has been modified slightly such that $b = c$, thus creating a non-obvious pseudo-merohedral indexing ambiguity which must be resolved before merging multiple data sets. The data are intended to be a realistic simulation of the radiation damage to a lysozyme-sized protein forming ∼5 µm crystals shot with a 6 µm beam.

The first three images of each data set were processed with *DIALS* (Winter *et al.*, 2018) *via* *xia*2 (Winter, 2010). No prior space-group or unit-cell information was provided, and integration was performed in *P*1.

The resulting 100 integrated data sets were analysed using the algorithms outlined in §2. A resolution cutoff of 3 Å was used; however, the results were not sensitive to the choice of resolution cutoff.

The 100 data sets had a median unit cell of $a = 38.31 \pm 0.03$, $b = 79.11 \pm 0.05$, $c = 79.12 \pm 0.07$ Å, $\alpha = 89.99 \pm 0.02$, $\beta = 89.99 \pm 0.03$, $\gamma = 90.00 \pm 0.01^\circ$. The maximum possible lattice symmetry was determined to be *P*422 (space group No. 89), comprising eight symmetry operations.

A bimodal distribution of r_{i_k,j_l} values can be seen in Fig. 1(*a*), which suggests the presence of an indexing ambiguity. Fig. 1(*b*) shows the resulting coordinates, **x**, projected onto the *xy* axes, and Fig. 1(*c*) shows the same coordinates projected onto the first two directions found by principal component analysis. The first direction identified by PCA accounts for 48% of the variance of the data, compared with only 11% for the second direction, and Fig. 1(*c*) shows that the points are clearly separated into two clusters, reflecting the two possible indexing choices. Two clusters were identified, each containing 400 points, corresponding to four points per data set. Analysis of each cluster according to §2.3 correctly identified the Patterson group as *P*222.

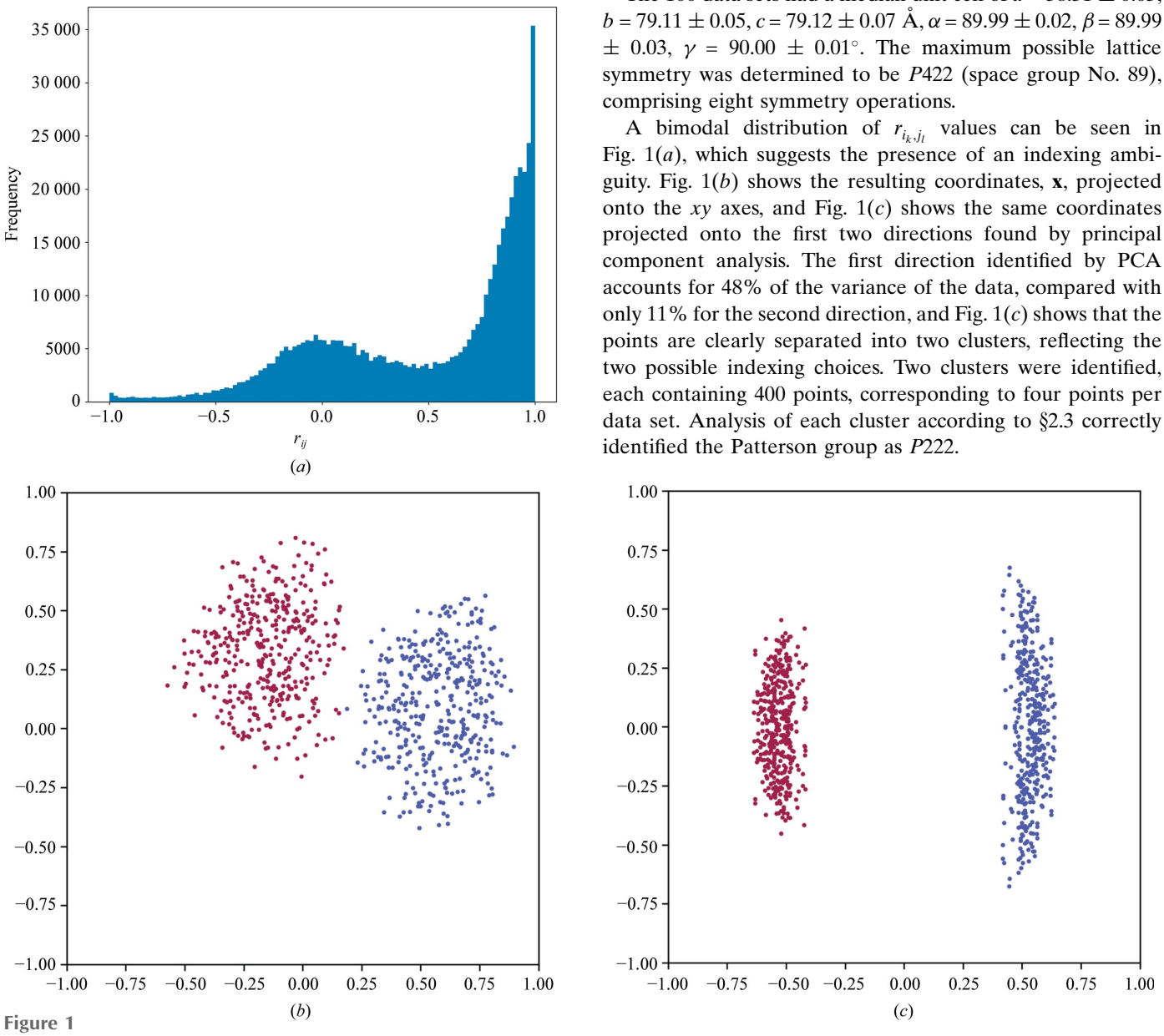


Figure 1
The application of the algorithms in §2 to simulated microfocus data sets as described in §3.1. A histogram of the r_{i_k,j_l} values is shown in (*a*). The points **x** determined by the procedure are shown projected onto the first two dimensions before (*b*) and after (*c*) principal component analysis. Points are coloured according to the assigned indexing mode.

3.2. Example 2: *in situ* membrane-protein data set

Previously published *in situ* data (Axford *et al.*, 2015) from an integral membrane protein, *Haemophilus influenzae* TehA (HiTehA), were reprocessed using *DIALS via xia*2. Processing was attempted on 72 wedges of data consisting of 30–50 images of 0.2° rotation, each wedge therefore consisting of 6–10° of data. No prior space-group or unit-cell information was provided, and integration was performed in *P*1 with the reduced unit cell. Two data sets failed in indexing, leaving 70 data sets which were subsequently analysed according to the algorithms described above.

The 70 data sets had a median unit cell of $a = 72.58 \pm 0.36$, $b = 72.74 \pm 0.29$, $c = 72.79 \pm 0.23$ Å, $\alpha = 85.16 \pm 0.08$, $\beta = 85.19 \pm 0.09$, $\gamma = 85.29 \pm 0.17$°. The maximum possible lattice symmetry was determined to be *R*32:r (space group No. 155), comprising six symmetry operations.

A bimodal distribution of $r_{i_k \cdot j_l}$ values can be seen in Fig. 2(*a*), which suggests the presence of an indexing ambiguity. The first direction identified by principal component analysis accounts for 67% of the variance of the data, compared with only 9.6% for the second direction, and visualization of the coordinates after projection onto the first two directions found by principal component analysis in

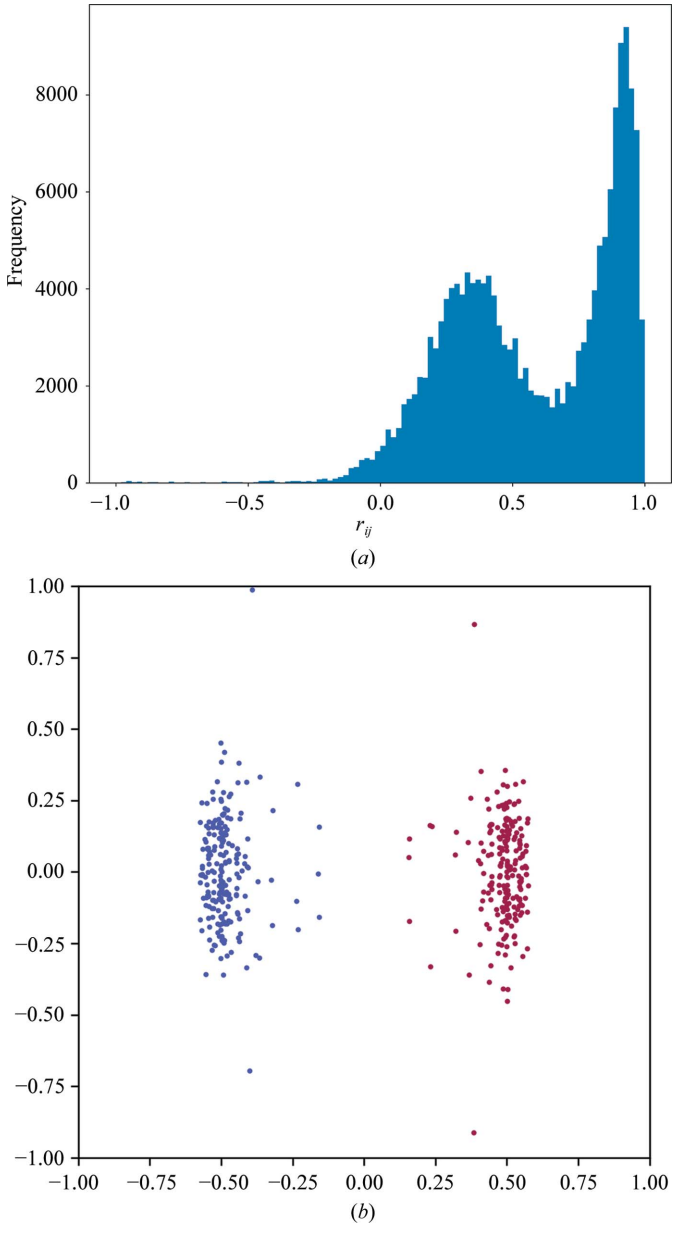


Figure 2
The application of the algorithms in §2 to the TehA multi-crystal data as described in §3.2. A histogram of the $r_{i_k \cdot j_l}$ values is shown in (*a*). The points **x** determined by the procedure are shown projected onto the first two dimensions identified by principal component analysis (*b*). Points are coloured according to the assigned indexing mode.

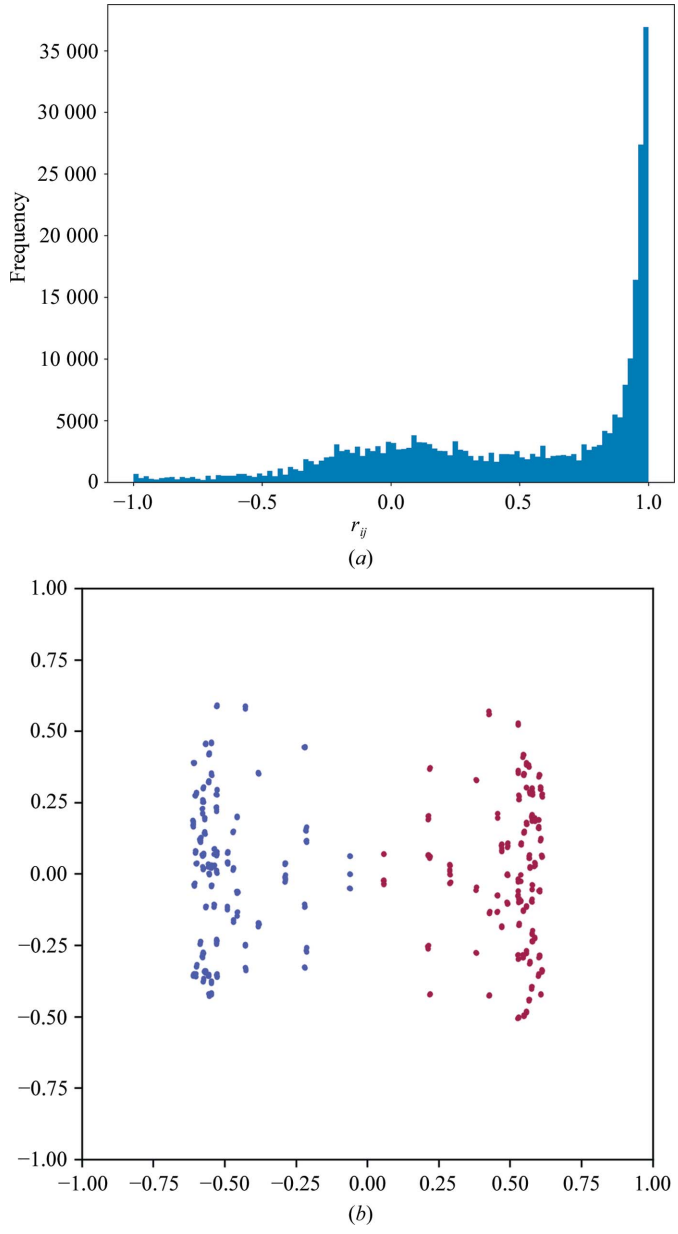


Figure 3
The application of the algorithms in §2 to multi-crystal data from CPV polyhedrin as described in §3.3. A histogram of the $r_{i_k \cdot j_l}$ values is shown in (*a*). The points **x** determined by the procedure are shown projected onto the first two dimensions identified by principal component analysis (*b*). Points are coloured according to the assigned indexing mode.

Fig. 2(*b*) shows that the points are clearly separated into two clusters, indicating the presence of two possible indexing modes. Two clusters were identified, each containing 210 points, corresponding to three points per data set. Analysis of each cluster according to §2.3 correctly identified the Patterson group as *R*3:h (space group No. 146), which is consistent with the published space group.

3.3. Example 3: *in cellulo* micro-crystal room-temperature data set

Forty 2° wedges of *in cellulo* data from cytoplasmic polyhedrosis virus (CPV) polyhedrin crystals (Axford *et al.*, 2014) were reprocessed using *DIALS via xia*2. No prior space-group or unit-cell information was provided, and integration was performed in *P*1 with the reduced unit cell. 28 data sets were successfully indexed and integrated, one of which was rejected after preliminary analysis using the hierarchical unit-cell clustering methods (Zeldin *et al.*, 2015) available within the *cctbx.xfel* software (Hattne *et al.*, 2014). The remaining 27 data sets had a median unit cell of $a = 88.92 \pm 0.17$, $b = 89.00 \pm 0.14$, $c = 89.04 \pm 0.12$ Å, $\alpha = 109.50 \pm 0.08$, $\beta = 109.44 \pm 0.09$, $\gamma = 109.38 \pm 0.08$°. The maximum possible lattice symmetry was determined to be *I*432 (space group No.211), comprising 24 symmetry operations.

A bimodal distribution of r_{i_k,j_l} values can be seen in Fig. 3(*a*), which suggests the presence of an indexing ambiguity. The first direction identified by principal component analysis accounts for 24% of the variance of the data, compared with only 6.2% for the second direction, and visualization of the coordinates after projection onto the first two directions found by principal component analysis in Fig. 3(*b*) shows a separation of the points into two clusters, indicating the presence of two possible indexing modes. Each of the two clusters identified contained 324 points, corresponding to 12 points per data set. Analysis of each cluster according to §2.3 correctly identified the Patterson group as *I*23 (space group No. 197), which is consistent with the published space group.

4. Discussion

The results shown in §3 demonstrate that it is possible to determine the Patterson group for sparse data sets in the presence of an indexing ambiguity. Three different examples were shown, covering simulated data sets with a pseudo-merohedral indexing ambiguity and previously published *in situ* and *in cellulo* multi-crystal data sets. In all cases, the data were reprocessed in space group *P*1 with no prior assumptions regarding the unit cell or symmetry. Application of the algorithms presented in §2 shows a separation of the resulting points into two clusters, representing the two alternative indexing choices. Further analysis of the composition of the clusters was able to correctly identify the correct Patterson group symmetry.

It is noteworthy that while the analysis defined above is predicated on the use of an *m*-dimensional space, where *m* is the number of symmetry operations in the lattice group, in many cases a lower dimensional analysis will give rise to a similar conclusion, particularly where the final number of clusters is small. In the above examples, the analysis was repeated with only two dimensions, resulting in the same conclusions.

Above, we refer to potential twinning operators as the sources of potential indexing ambiguity. The presence of partial twinning would have the effect of making the intensities of the alternative indexing possibilities more similar, thus reducing the separation between the peaks in the histogram of r_{i_k,j_l} values. This would be expected to reduce the angular separation between the clusters of points, $\mathbf{x}_i$, output by the algorithm. As such, we expect our algorithm to be tolerant to the presence of partial twinning when the twin fraction is small, albeit with reduced sensitivity. However, the power of the algorithm to distinguish between indexing modes will rapidly reduce as the twin fraction approaches that for perfect twinning, *i.e.* $\alpha = 0.5$.

Once any potential symmetry and indexing ambiguities have been identified and resolved, existing methods for the determination of the space group (Evans, 2006) and clustering of data sets based on unit-cell parameters (Foadi *et al.*, 2013) and intensities (Giordano *et al.*, 2012; Diederichs, 2017; Santoni *et al.*, 2017) may be used. The algorithms presented here allow data to be integrated in *P*1 with no prior assumptions, with conclusions relating to symmetry derived from the data set as a whole. They therefore make a useful addition to the tools for *in situ* data processing.

Acknowledgements

The authors would like to thank Danny Axford for provision of the raw data from the TehA and CPV experiments, and James Holton for generating and making available the microfocus data-processing challenge images, as well as the MX team at Diamond Light Source and the wider *DIALS* collaboration. The tools were developed using the *cctbx* and *DIALS* toolkits and will be included within *DIALS* and *xia*2.

Funding information

This work was funded by Diamond Light Source.

References

Axford, D., Foadi, J., Hu, N.-J., Choudhury, H. G., Iwata, S., Beis, K., Evans, G. & Alguel, Y. (2015). *Acta Cryst.* D**71**, 1228–1237.
Axford, D., Ji, X., Stuart, D. I. & Sutton, G. (2014). *Acta Cryst.* D**70**, 1435–1441.
Brehm, W. & Diederichs, K. (2014). *Acta Cryst.* D**70**, 101–109.
Diederichs, K. (2017). *Acta Cryst.* D**73**, 286–293.
Evans, P. (2006). *Acta Cryst.* D**62**, 72–82.
Flack, H. D. (1987). *Acta Cryst.* A**43**, 564–568.
Foadi, J., Aller, P., Alguel, Y., Cameron, A., Axford, D., Owen, R. L., Armour, W., Waterman, D. G., Iwata, S. & Evans, G. (2013). *Acta Cryst.* D**69**, 1617–1632.
Gildea, R. J., Waterman, D. G., Parkhurst, J. M., Axford, D., Sutton, G., Stuart, D. I., Sauter, N. K., Evans, G. & Winter, G. (2014). *Acta Cryst.* D**70**, 2652–2666.

Ginn, H. M., Messerschmidt, M., Ji, X., Zhang, H., Axford, D., Gildea, R. J., Winter, G., Brewster, A. S., Hattne, J., Wagner, A., Grimes, J. M., Evans, G., Sauter, N. K., Sutton, G. & Stuart, D. I. (2015). *Nature Commun.* **6**, 6435.
Giordano, R., Leal, R. M. F., Bourenkov, G. P., McSweeney, S. & Popov, A. N. (2012). *Acta Cryst.* D**68**, 649–658.
Grosse-Kunstleve, R. W., Sauter, N. K. & Adams, P. D. (2004). *IUCr Comm. Crystallogr. Comput. Newsl.* **3**, 22–31.
Hattne, J. *et al.* (2014). *Nature Methods*, **11**, 545–548.
Holton, J. (2015). *The Micro-focus Data Processing Challenge.* http://bl831.als.lbl.gov/~jamesh/challenge/microfocus/.
Kabsch, W. (2014). *Acta Cryst.* D**70**, 2204–2216.
Lebedev, A. A., Vagin, A. A. & Murshudov, G. N. (2006). *Acta Cryst.* D**62**, 83–95.
Le Page, Y. (1982). *J. Appl. Cryst.* **15**, 255–259.
Liu, D. C. & Nocedal, J. (1989). *Math. Program.* **45**, 503–528.
Mayans, O., Wuerges, J., Canela, S., Gautel, M. & Wilmanns, M. (2001). *Structure*, **9**, 331–340.
Pearson, K. (1901). *Lond. Edinb. Dubl. Philos. Mag. J. Sci.* **2**, 559–572.
Santoni, G., Zander, U., Mueller-Dieckmann, C., Leonard, G. & Popov, A. (2017). *J. Appl. Cryst.* **50**, 1844–1851.
Sauter, N. K., Grosse-Kunstleve, R. W. & Adams, P. D. (2006). *J. Appl. Cryst.* **39**, 158–168.
White, T. A., Mariani, V., Brehm, W., Yefanov, O., Barty, A., Beyerlein, K. R., Chervinskii, F., Galli, L., Gati, C., Nakane, T., Tolstikova, A., Yamashita, K., Yoon, C. H., Diederichs, K. & Chapman, H. N. (2016). *J. Appl. Cryst.* **49**, 680–689.
Winter, G. (2010). *J. Appl. Cryst.* **43**, 186–190.
Winter, G., Waterman, D. G., Parkhurst, J. M., Brewster, A. S., Gildea, R. J., Gerstel, M., Fuentes-Montero, L., Vollmar, M., Michels-Clark, T., Young, I. D., Sauter, N. K. & Evans, G. (2018). *Acta Cryst.* D**74**, 85–97.
Zeldin, O. B., Brewster, A. S., Hattne, J., Uervirojnangkoorn, M., Lyubimov, A. Y., Zhou, Q., Zhao, M., Weis, W. I., Sauter, N. K. & Brunger, A. T. (2015). *Acta Cryst.* D**71**, 352–356.