PROCEEDINGS PAPER

Bi-cross validation of spectral clustering hyperparameters

Sioan Zohar,[a] and Chun Hong Yoon
Photon Data and Controls Systems, Linac Coherent Light Source, SLAC National Accelerator Laboratory, 2575 Sand Hill Rd, Menlo Park, California 94025, USA

(Received 15 March 2020; accepted 23 March 2020)

One challenge impeding the analysis of terabyte scale X-ray scattering data from the Linac Coherent Light Source (LCLS) is determining the number of clusters required for the execution of traditional clustering algorithms. Here, we demonstrate that the previous work using bi-cross validation to determine the number of singular vectors directly maps to the spectral clustering problem of estimating both the number of clusters and hyperparameter values. Applying this method to LCLS X-ray scattering data enables the identification of dropped shots without manually setting boundaries on detector fluence and provides a path toward identifying rare and anomalous events. © *International Centre for Diffraction Data 2020*. [doi:10.1017/S0885715620000214]

Key words: XRD, spectroscopy, machine learning, *in situ* diffraction

I. INTRODUCTION

X-ray free electron lasers (X-FELs) (Ishikawa *et al*., 2012) are remarkable instruments capable of producing highly coherent X-ray pulses less than 20 fs in duration. Since their inception, X-FELs have made contributions to a diverse range of disciplines spanning from condensed matter (Higley *et al*., 2019) and atomic molecular optics (Yang *et al*., 2018) to structural biology (Nogly *et al.*, 2018) and femtosecond chemistry (Hong *et al.*, 2015). Compared to third-generation light sources, X-FELs require high-throughput data systems (Thayer *et al*., 2016) for writing to the disk on a per-pulse basis. Originally developed in order to filter out low fluence shots in post processing, shot-by-shot recording has since shifted the data collection paradigm and provided researchers with the means to compensate X-ray/laser timing jitter (Droste *et al*., 2019), outrun X-ray damage accumulation in protein crystallography experiments (Kupitz *et al*., 2017; Spence, 2017), and offers the potential to extract new physics by identifying rare events (Schoenlein *et al*., 2017).

Data accumulated over the course of a Linac Coherent Light Source (LCLS) user experiment regularly exceeds 20 TB and approximately 2.5 years analyzing such data is required before the results are published. Efforts to expedite the analysis have motivated the development of a high-performance computing infrastructure, novel algorithms (Yoon *et al.*, 2011), and user-friendly abstraction layers (Damiani *et al.*, 2016) similar to graphical user interfaces provided by commercial software vendors. One promising avenue for streamlining data analysis is the exploitation of clustering algorithms. Such algorithms are currently used to cluster diffraction images of protein conformations collected in diffract and destroy experiments (Yoon *et al.*, 2011). With the increased data rates anticipated for LCLS2, clustering algorithms will have the potential to identify and isolate the rare events of charge separation, migration, and accumulation during multi-step catalytic processes in molecular complexes and devices (Schoenlein *et al*., 2017). One impediment to achieving these goals is the challenge of estimating the hyperparameters and the number of clusters required for the execution of clustering algorithms.

k-Means clustering is the process of labeling data based solely on the distribution of the data itself. For linearly separable clusters, this is accomplished by drawing a set of decision boundaries in the form of hyperplanes that minimize the intracluster variance summed over all clusters (Lloyd, 1982). This method, however, prescribes no approach for how many clusters one should choose. Early works estimating the number of clusters used a combination of gap methods (Tibshirani *et al*., 2001), distortion methods (Sugar and James, 2003), stability approaches (Tibshirani and Walther, 2005; Von Luxburg, 2010), and nonparametric methods (Fujita *et al*., 2014). These approaches are generally considered to be heuristic with well-understood limitations and require assumptions about the cluster distribution. More recent work (Fu and Perry, 2019) has made exciting progress in both implementing and laying the theoretical foundation for abstracting bi-cross validation (BCV) (Owen and Perry, 2009) away from its matrix formulation to estimate the number of clusters for use with the *k*-means algorithm. This approach, however, requires preconditioning rotations to discriminate when multiple clusters are spaced along a single feature dimension and can only label clusters that are linearly separable. In that work (Fu and Perry, 2019), it was predicted that applying BCV to the Laplacian matrix after the eigenvector transformation would provide a convex loss function for estimating the number of clusters.

Here, it is shown that spectral clustering hyperparameters, including the number of clusters, can be estimated by performing BCV on the inverted Laplacian matrix and finding the local minima of the resultant BCV loss function. In spectral clustering, data are embedded into a higher dimensional graph representation called the Laplacian matrix (Von Luxburg, 2007, 2010). The multiplicity of the Laplacian's smallest eigenvalues is equal to the number of clusters.

[a]Author to whom correspondence should be addressed. Electronic mail: zohar.sioan@gmail.com

 0885-7156/2020/35(2)/112/5/$18.00

BCV is a powerful least squares method for estimating the number of dominant singular vectors needed to reconstruct the matrix without over fitting the data to the noise (Owen and Perry, 2009). Inverting the Laplacian matrix converts the problem of cluster number estimation from one of estimating the number of smallest singular vectors into the problem of estimating the number of largest singular vectors that, in turn, can be solved using BCV.

The main result of this work is captured in Eq. (7) which connects the spectral clustering and BCV frameworks. The range where this technique succeeds and fails is explored using simulated data sets. Applying this technique to experimental LCLS X-ray scattering data separates low fluence from high fluence X-ray pulses and provides a path toward identifying clusters of rare events.

II. THEORY

We consider a set of X-ray scattering data stored within a matrix $\mathbf{X}$, with elements $\mathbf{X_{ij}}$ where $\mathbf{i}$ and $\mathbf{j}$ are the row and column indices, respectively. All entries contained within a row were measured at the same instant, and all entries within a single column measure the same quantity. For the case of LCLS data, potential column labels are incident X-ray pulse energy, scattered pulse energy, photon energy, X-ray/laser jitter correction, or laser delay stage position. The process of clustering, in this context, means creating columns that assign labels such as "signal of interest", "low fluence shots", "outliers", or "rare events".

In the spectral clustering approach, clusters are identified by applying k-means clustering on the k smallest eigenvectors, v, of the Laplacian matrix, $\mathbf{L}$, where k is the number of clusters. Formally,

$$\mathbf{L} = \mathbf{W} - \mathbf{D} \tag{1}$$

where $\mathbf{D}$ is the degree matrix. The weighting $\mathbf{W}$ matrix chosen here is calculated using the radial basis function (RBF) kernel (Chung *et al.*, 2003) such that

$$\mathbf{W_{i,j}} = \exp\left[-\sum_{\mathbf{m}} (\mathbf{X_{i,m}} - \mathbf{X_{j,m}})^{\mathbf{T}} \mathbf{\Gamma} (\mathbf{X_{i,m}} - \mathbf{X_{j,m}})\right] \tag{2}$$

where $\mathbf{i}$ and $\mathbf{j}$ are the row and column indices of $\mathbf{W}$, and $\mathbf{\Gamma}$ is a hyperparameter that is inversely proportional to the root of the expected distance between points within a cluster. Traditionally, $\mathbf{\Gamma}$ is treated as a scalar. In practice, the Laplacian is normalized by

$$\mathbf{L_n} = \mathbf{D^{-1/2}}\,\mathbf{L}\,\mathbf{D^{-1/2}} = \mathbf{I} - \mathbf{D^{-1/2}}\mathbf{W}\,\mathbf{D^{-1/2}} \tag{3}$$

where $\mathbf{L_n}$ is the normalized Laplacian. Using these definitions, the spectral clustering method proceeds by solving the generalized eigenvector problem

$$\mathbf{L_n} v = \lambda \mathbf{D} v, \tag{4}$$

implementing k-means on the diagonalized feature space and propagating the resultant labels from k-means back to $\mathbf{X}$.

The procedure for estimating the number of clusters and $\mathbf{\Gamma}$ by performing BCV on the inverted Laplacian, $\mathbf{L}^{-1}$, proceeds as follows. The Laplacian is, by construction, a singular matrix that cannot be inverted. This drawback is circumvented by introducing a regularization term, $\mathbf{R}$. That is

$$\mathbf{L_r} = \mathbf{L_n} + \xi \mathbf{R} \tag{5}$$

where ξ is a scalar regularization parameter. Here, ξ is empirically determined to be of the order 1×10^{-9} to 1×10^{-14}. The matrix $\mathbf{R}$ is

$$\mathbf{R} = \mathbf{H} - \mathbf{H^T}\mathbf{L_n}\mathbf{H} \tag{6}$$

where $\mathbf{H}$ is a Haar distributed random matrix (Mezzadri, 2006). Adding $\xi\,\mathbf{R}$ to $\mathbf{L_n}$, as opposed to adding $\xi\,\mathbf{H}$ directly, guarantees that the resultant matrix $\mathbf{L_r}$ can be inverted. The BCV loss function for $\mathbf{L_r^{-1}}$ is calculated as described in Owen and Perry (2009) by breaking $\mathbf{L_r^{-1}}$ into quadrants.

$$\mathbf{L_r^{-1}} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{E} \end{bmatrix} \tag{7}$$

The bottom-right quadrant has been labeled in this work as $\mathbf{E}$, deviating from the notation in the previous literature (Owen and Perry, 2009) so as not to be confused with the degree matrix $\mathbf{D}$. Here, $\mathbf{A}$ was designated as the holdout, and 2×2 BCV was configured such that the sub-matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{E}$ have the same number of rows and columns. This submatrix partitioning is close to the optimal 52% holdout size for square matrices (Perry, 2009). The BCV loss function is

$$\mathrm{BCV}(k, \mathbf{\Gamma}) = \sum_{\mathbf{i,j}} (\mathbf{A} - \mathbf{B}(\hat{\mathbf{E}}^k)^{+}\mathbf{C})^2_{\mathbf{i,j}} \tag{8}$$

where $(\hat{\mathbf{E}}^k)^+$ is the Penrose pseudo inverse of $(\hat{\mathbf{E}}^k)$,

$$(\hat{\mathbf{E}}^k)^{+} = ((\hat{\mathbf{E}}^k)^{\mathbf{T}}\hat{\mathbf{E}}^k)^{-1}(\hat{\mathbf{E}}^k)^{\mathbf{T}} \tag{9}$$

and $\hat{\mathbf{E}}^k$ is the singular value decomposition (SVD) reconstruction of $\mathbf{E}$ using the k number of basis vectors. The procedure starting from Eq. (7) was iterated ~40 times with $\mathbf{L_r^{-1}}$ being shuffled each iteration before being decomposed into the submatrices in Eq. (7). The BCV score used to determine the number of clusters is the average BCV score over all iterations.

III. NUMERICAL SIMULATIONS

The performance of this approach was benchmarked for a range of hyperparameters using Scikit-learn version 0.19.1, Numpy version 1.14.2, and Scipy version 0.19.1 packages (Oliphant, 2006, 2007; Pedregosa *et al.*, 2011; Van Der Walt *et al.*, 2011). Source code containing an executable step-by-step walk through can be cloned from this repository (Zohar, 2019).

In Figure 1(a), a set of five simulated clusters projected from a seven-dimensional feature space onto two dimensions (2D) are shown. Panel (b) shows seven simulated clusters generated in a two-dimensional feature space. The BCV loss function minimum was found by iterating over increasing values of cluster number, k, and length scales, $\mathbf{\Gamma}$, and calculating the BCV loss function at each point. The BCV score's dependence on k for the clusters in panels (a) and (b) is shown in panels (c) and (d), respectively. The different color lines shown in panels (c) and (d) correspond to increasing values of the regularization parameter. The BCV score in panel (c)

https://doi.org/10.1017/S0885715620000214 Published online by Cambridge University Press

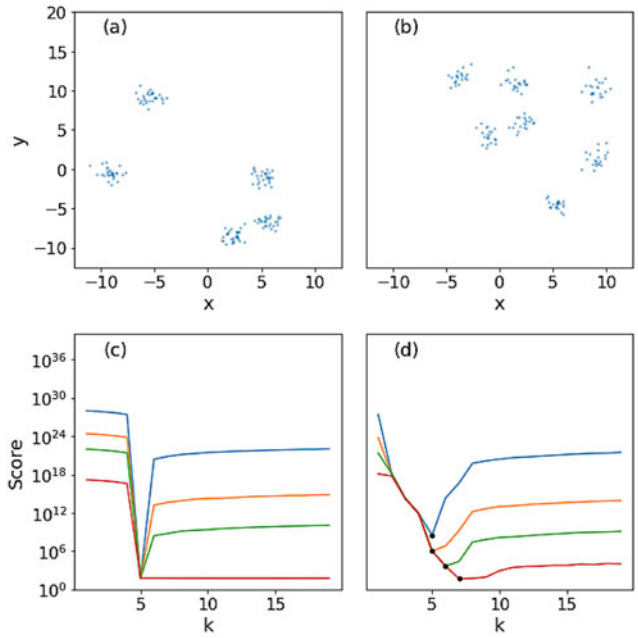

Figure 1. (Color online) (a) A set of 150 samples occupying a 7D feature space are clustered into five groups and projected onto 2D. (b) The intercluster spacing is reduced by reducing the feature space from 7D to 2D and increasing the number of clusters from 5 to 7. (c) The BCV score dependence on the number of clusters for regularization parameter values of 1×10^{-14} (blue), 6.3×10^{-13} (orange), 1×10^{-12} (green), and 2.5×10^{-9} (red). The score minimum occurs at 5 which is the expected number of clusters. (d) When the interclustering spacing is reduced, BCV does not robustly estimate the number of clusters, since the score minimum (black dots) does not occur at the same k value for all values of ξ and only occurs at the expected value of 7 for $\xi=2.5\times10^{-9}$.

has a minimum at $k=5$ correctly identifying the number of clusters. This estimate is robust for changing regularization values except for large regularization, where the score minimum no longer occurs at the expected number of clusters and moves to arbitrarily large k. The intercluster distance for points in panel (b) is decreased with respect to panel (a) by increasing the number of clusters from 5 to 7 and reducing the feature space dimension from 7 to 2. The BCV score for the points in panel (b) is shown in panel (d). For a fixed value of $\mathbf{\Gamma}$, the cluster number estimation procedure is not robust since the score minimum does not reliably estimate the number of clusters for all values of ξ.

In Figure 2(a), a set of clusters in 2D are shown. The clusters can be partitioned into 3 or 11 different groups, depending on the Gaussian kernel width, $\mathbf{\Gamma}$, chosen to construct the affinity matrix. The BCV scores plotted as a function of the number of clusters are shown in panels (b), (c), (d), and (e) for values of $\mathbf{\Gamma}$ equal to 0.005, 0.028, 0.158, and 1.58, respectively. The different colored curves are for different values of the regularization parameter ξ. For the smallest regularization values (blue curves), two global minima occurring at $\mathbf{\Gamma}$ values of 1.58 [Figure 2(b)] and 0.005 [Figure 2(e)] occur at k equal to 3 and 11, respectively. In Figure 3, a heat map of the BCV's score value's dependence upon the Gaussian kernel width and the number of clusters is shown for $\xi=10^{-14}$. The RBF parameter $\mathbf{\Gamma}$ can be converted into a characteristic length scale σ, using $\mathbf{\Gamma}=1/(2\sigma^2)$. The two local minima observed at $k=11$ and $k=3$ have corresponding σ values of the orders of 1 and 10, respectively, which correspond to two different length scales at which the clusters can be grouped. The ability to estimate both the number of clusters and the spectral clustering $\mathbf{\Gamma}$ hyperparameter is advantageous compared to previous methods which provide a loss function that estimates the number of clusters but not any additional hyperparameters.

IV. EXPERIMENTAL DATA

In Figure 4, the results from applying this approach to an X-ray scattering experiment are shown. The scattered X-ray intensity, incident X-ray intensity, photon energy, and other machine parameters were measured at the soft X-ray beamline at the SLAC Linear Accelerator's LCLS just below the Cu L_3 edge. The sample under study was an yttrium barium copper oxide (YBCO) thin film that has been shown to exhibit high-temperature superconductivity. The feature space is 12 dimensions with column labels corresponding to the intensity of

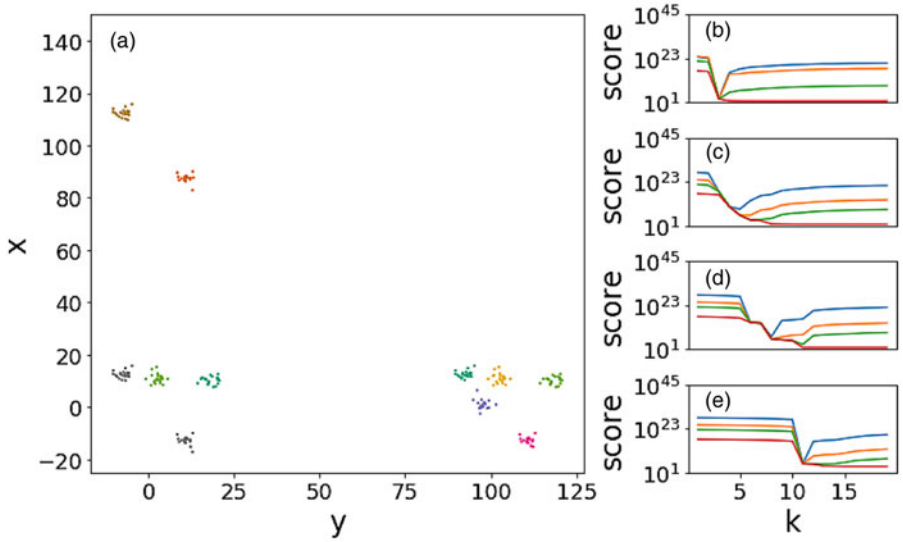

Figure 2. (Color online) Demonstration of cluster identification at different length scales. (a) A set of 150 samples clustered into 11 groups that appear as three clusters on longer length scales. (b) Density map of their score dependence on $\mathbf{\Gamma}$ and k. Regularization values are 1×10^{-14} (blue), 6.3×10^{-13} (orange), 1×10^{-12} (green), and 2.5×10^{9} (red). The score as the function of the cluster number k is shown for $\mathbf{\Gamma}$ equal to 0.005, 0.028, 0.158, and 1.58 for panels (b), (c), (d), and (e), respectively.

https://doi.org/10.1017/S0885715620000214 Published online by Cambridge University Press

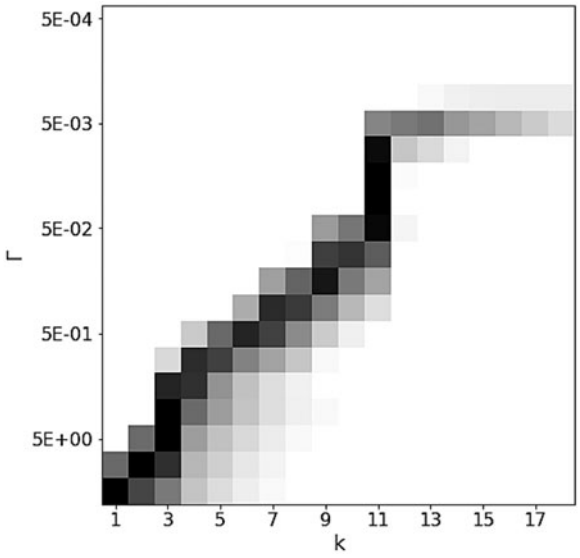

Figure 3. Density map of the score dependence on Γ and k for $\xi = 1 \times 10^{-14}$. The dark and light regions correspond to low and high BCV loss function values, respectively. Cross sections of this density map for fixed values of Γ are shown in Figure 2(b)–(e).

X-rays scattered off the sample, the incident intensity downstream of the monochromator, four different incident intensity diagnostics from upstream of the monochromator, laser delay-stage position, laser power, arrival time monitor mean and FWHM, photon energy, and the photon energy product with the incident intensity downstream of the monochromator. Multiplying the photon energy with the intensity linearizes the chromatic nonlinearity observed when the photon energy is tuned to the steep part of an X-ray absorption edge (Zohar and Turner, 2019).

The problem of heterogeneous density present in spectral clustering is circumvented by feature engineering an additional column that contains an estimate of the point density in the local vicinity. This was accomplished by appending the diagonal values of the degree matrix, calculated for 7000 samples using a $\mathbf{\Gamma} = 1 \times 10^{-2}$, to the feature space. Clustering was performed on a total of 750 rows from this feature space. Eleven clusters are identified with the populations of the dominant first three clusters containing on average 573, 62, and 43 data points. The rest of the data points are spread over the remaining clusters. As shown in Figure 4, this approach separates out the dominant cluster (blue histogram), which corresponds to signal of interest from the dropped shots with no fluence (orange histogram). It is stressed that the last figure presented here is analyzed on less than 1% of the entire data and does not represent the expected number of clusters if the full data sets were to be used.

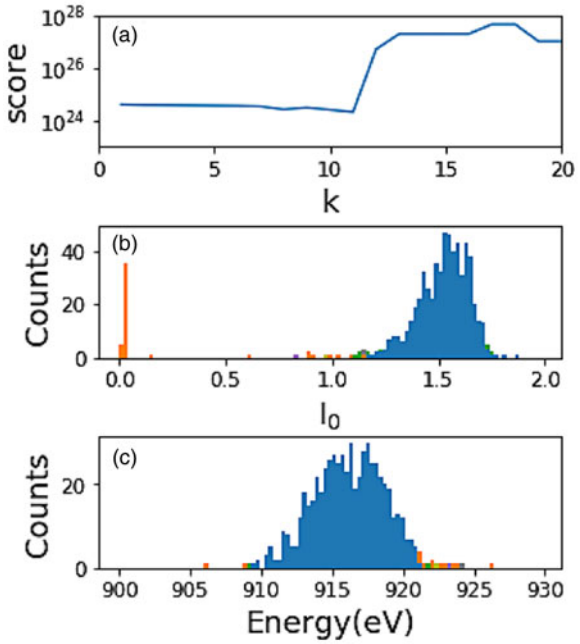

Figure 4. (Color online) (a) The BCV score minimum occurs for 11 clusters. (b) Histogram of the incident pulse energy measured in a gas detector upstream of the monochromator. The orange and blue histograms correspond to the dropped shots and signal of interest, respectively. (c) Histogram of the photon energy generated upstream of the monochromator.

V. DISCUSSION

There are several advantages for using the matrix formulation (Owen and Perry, 2009) of BCV as opposed to the abstracted BCV form in the nonembedded feature space (Fu and Perry, 2019). One advantage is that the preconditioning rotation steps needed for preventing clusters from laying along one non-separable dimension are no longer required. Another advantage is that since the matrix BCV formulation does not require a classification step, there are no additional hyperparameters that need to be estimated.

The ability to simultaneously estimate both the $\mathbf{\Gamma}$ parameter and the number of clusters is not serendipitous. Intuitively, it is understood that asking "how many clusters are present in some region" cannot be separated from the question of "what length scales do those same clusters appear on?". This line of thinking agrees with the limiting cases of very small and very large $\mathbf{\Gamma}$ values, where the number of estimated clusters will be equal to either one or the number of points respectively. Looking forward, there are several prerequisites that would need to be met for this approach to be widely adopted. A mathematical proof demonstrating that the BCV loss function minimum correctly estimates the hyperparameters would have to be shown. This proof would provide insight on how to estimate the regularization parameter by exploiting the regularized Laplacian's condition number and using the eigenvector decomposition of the inverted Laplacian as opposed to SVD decomposition.

VI. CONCLUSION

A direct matrix implementation of BCV for estimating both the number of clusters and kernel hyperparameters used in spectral clustering has been demonstrated. This was accomplished by applying the matrix formulation for BCV directly to the inverted Laplacian matrix. The resulting BCV loss function has robust minima that occur at different cluster numbers depending upon the length scales determined by the RBF kernel parameter. The results here provide a path toward generalized hyperparameter optimization for spectral clustering algorithms.

ACKNOWLEDGEMENTS

We thank Art B. Owen for providing fruitful discussions and insights. This work was performed in support of the LCLS project at SLAC supported by the U.S. Department of Energy, Office of Science, and Office of Basic Energy Sciences, under Contract No. DE-AC02-76SF00515.

https://doi.org/10.1017/S0885715620000214 Published online by Cambridge University Press

Chung, K.-M., Kao, W.-C., Sun, C.-L., Wang, L.-L., and Lin, C.-J. (**2003**). "Radius margin bounds for support vector machines with the RBF kernel," Neural Comput. **15**, 2643.

Damiani, D., Dubrovin, M., Gaponeneko, I., Kroeger, W., Lane, T., Mitra, A., O'Grady, C., Salnikov, A., SanchezGonzalez, A., Schneider, D. *et al*. (**2016**). "Linac Coherent Light Source data analysis using psana," J. Appl. Crystallogr. **49**, 672.

Droste, S., Shen, L., White, V. E., Diaz-Jacobo, E., Coffee, R., Zohar, S., Reid, A. H., Tavella, F., Minitti, M. P., Turner, J. J., Gumerlock, K. L., Fry, A. R., and Coslovich, G. (**2019**). "High-sensitivity X-ray optical cross-correlator for next generation free-electron lasers," *CLEO: OSA Technical Digest (Optical Society of America, 2019),* pp. SF3I–7. https://www.osapublishing.org/abstract.cfm?uri=CLEO_SI-2019-SF3I.7

Fu, W. and Perry, P. O. (**2019**). ""Estimating the number of clusters using cross-validation." J. Comput. Graph. Stat., 1–12.

Fujita, A., Takahashi, D. Y., and Patriota, A. G. (**2014**). "A non-parametric method to estimate the number of clusters," Comput. Stat. Data Anal. **73**, 27.

Higley, D. J., Reid, A. H., Chen, Z., Guyader, L. L., Hellwig, O., Lutman, A. A., Liu, T., Shafer, P., Chase, T., Dakovski, G. L., Mitra, A., Yuan, E., Schlappa, J., Durr, H. A., Schlotter, W. F., and Stohr, J. (**2019**). "Ultrafast X-ray induced changes of the electronic and magnetic response of solids due to valence electron redistribution." Preprint, arXiv:1902.04611.

Hong, K., Cho, H., Schoenlein, R. W., Kim, T. K., and Huse, N. (**2015**). "Element-specific characterization of transient electronic structure of solvated Fe (II) complexes with time-resolved soft X-ray absorption spectroscopy," Acc. Chem. Res. **48**, 2957.

Ishikawa, T., Aoyagi, H., Asaka, T., Asano, Y., Azumi, N., Bizen, T., Ego, H., Fukami, K., Fukui, T., Furukawa, Y. *et al*. (**2012**). "A compact X-ray free-electron laser emitting in the sub-ångström region," Nature Phonotonics **6**, 540C.

Kupitz, C., Olmos, J. L. Jr., Holl, M., Tremblay, L., Pande, K., Pandey, S., Oberthür, D., Hunter, M., Liang, M., Aquila, A. *et al*. (**2017**). "Structural enzymology using X-ray free electron lasers," Structural Dynamics **4**, 044003.

Lloyd, S. P. (**1982**). "Least squares quantiation in PCM," IEEE Trans. Inf. Theory **28**, 129.

Mezzadri, F. (**2006**). "How to generate random matrices from the classical compact groups," Notices Am. Math. Soc. **54**, 592.

Nogly, P., Weinert, T., James, D., Carbajo, S., Ozerov, D., Furrer, A., Gashi, D., Borin, V., Skopintsev, P., Jaeger, K. *et al*. (**2018**). "Retinal isomerization in bacteriorhodopsin captured by a femtosecond x-ray laser," Science **361**, eaat0094.

Oliphant, T. E. (**2006**). *A Guide to NumPy*, Vol. 1 (Trelgol Publishing, USA).

Oliphant, T. E. (**2007**). "Python for scientific computing," Comput. Sci. Eng. **9**, 10.

Owen, A. B. and Perry, P. O. (**2009**). "Bi-cross-validation of the SVD and the nonnegative matrix factorization," Ann. Appl. Stat. **3**, 564.

Pedregosa, F. *et al*. (**2011**). "Scikit-learn: machine learning in python," J. Mach. Learn. Res. **12**, 2825.

Perry, P. O. (**2009**). "Cross-validation for unsupervised learning." Preprint, arXiv:0909.3052.

Schoenlein, R., Boutet, S., Minitti, M., and Dunne, A. (**2017**). "The Linac Coherent Light Source: recent developments and future plans," Appl. Sci. **7**, 850.

Spence, J. C. (**2017**). "Outrunning damage: electrons vs X-rays – timescales and mechanisms," Struct. Dyn. **4**, 044027.

Sugar, C. A. and James, G. M. (**2003**). "Finding the number of clusters in a dataset," J. Am. Stat. Assoc. **98**, 750.

Thayer, J., Damiani, D., Ford, C., Gaponenko, I., Kroeger, W., O'Grady, C., Pines, J., Tookey, T., Weaver, M., and Perazzo, A. (**2016**). "Data systems for the Linac Coherent Light Source," J. Appl. Crystallogr. **49**, 1363–1369.

Tibshirani, R. and Walther, G. (**2005**). "Cluster validation by prediction strength," J. Comput. Graph. Stat. **14**, 511.

Tibshirani, R., Walther, G., and Hastie, T. (**2001**). "Estimating the number of clusters in a data set via the gap statistic," J. R. Stat. Soc. Ser. B Stat. Methodol. **63**, 411.

Van Der Walt, S., Colbert, S. C., and Varoquaux, G. (**2011**). "The NumPy array: a structure for efficient numerical computation," Comput. Sci. Eng. **13**, 22.

Von Luxburg, U. (**2007**). "A tutorial on spectral clustering," Stat. Comput. **17**, 395.

Von Luxburg, U. (**2010**). "Clustering stability: an overview," Found. Trends Mach. Learn. **2**, 235.

Yang, J., Zhu, X., Wolf, T. J., Li, Z., Nunes, J. P. F., Coffee, R., Cryan, J. P., Gühr, M., Hegazy, K., and Heinz, T. F. *et al*. (**2018**). "Imaging CF3I conical intersection and photodissociation dynamics with ultrafast electron diffraction," Science **361**, 64.

Yoon, C. H., Schwander, P., Abergel, C., Andersson, I., Andreasson, J., Aquila, A., Bajt, S., Barthelmess, M., Barty, A., and Bogan, M. J., *et al*. (**2011**). "Unsupervised classification of single-particle X-ray diffraction snapshots by spectral clustering," Opt. Express **19**, 16542.

Zohar, S. (**2019**) Available at: https://github.com/sioan/bcv_spectral_clustering/blob/master/spectral_clustering.ipynb (accessed 2019-05-7).

Zohar, S. and Turner, J. J. (**2019**). "Multivariate analysis of x-ray scattering using a stochastic source," Opt. Lett. **44**, 243.

https://doi.org/10.1017/S0885715620000214 Published online by Cambridge University Press