Testing for Multivariate Normality in Mass Spectrometry Imaging Data: A Robust Statistical Approach for Clustering Evaluation and the Generation of Synthetic Mass Spectrometry Imaging Data Sets

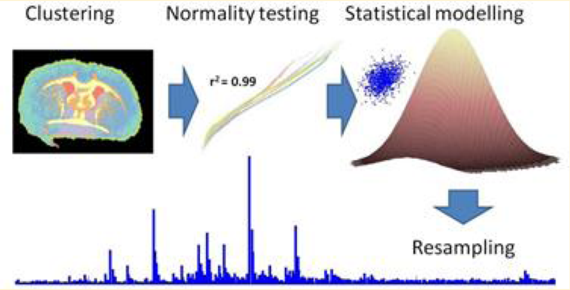
Alex Dexter,[†,‡] Alan M. Race,[‡] Iain B. Styles,[§] and Josephine Bunch*[,‡,∥]

[†]PSIBS Doctoral Training Centre, [§]School of Computer Science, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

[‡]National Physical Laboratory, Teddington, Middlesex TW11 0LW, United Kingdom

[∥]School of Pharmacy, University of Nottingham, Nottingham, Nottinghamshire NG7 2RD, United Kingdom

S *Supporting Information*



ABSTRACT: Spatial clustering is a powerful tool in mass spectrometry imaging (MSI) and has been demonstrated to be capable of differentiating tumor types, visualizing intratumor heterogeneity, and segmenting anatomical structures. Several clustering methods have been applied to mass spectrometry imaging data, but a principled comparison and evaluation of different clustering techniques presents a significant challenge. We propose that testing whether the data has a multivariate normal distribution within clusters can be used to evaluate the performance when using algorithms that assume normality in the data, such as *k*-means clustering. In cases where clustering has been performed using the cosine distance, conversion of the data to polar coordinates prior to normality testing should be performed to ensure normality is tested in the correct coordinate system. In addition to these evaluations of internal consistency, we demonstrate that the multivariate normal distribution can then be used as a basis for statistical modeling of MSI data. This allows the generation of synthetic MSI data sets with known ground truth, providing a means of external clustering evaluation. To demonstrate this, reference data from seven anatomical regions of an MSI image of a coronal section of mouse brain were modeled. From this, a set of synthetic data based on this model was generated. Results of r^2 fitting of the chi-squared quantile−quantile plots on the seven anatomical regions confirmed that the data acquired from each spatial region was found to be closer to normally distributed in polar space than in Euclidean. Finally, principal component analysis was applied to a single data set that included synthetic and real data. No significant differences were found between the two data types, indicating the suitability of these methods for generating realistic synthetic data.

Data mining is a valuable tool in mass spectrometry imaging (MSI), where even a single image can contain more information than can be feasibly interpreted by a single person in a realistic time frame. Often, a few m/z values or pixels of interest are selected for analysis based on known information about the sample. It is becoming increasingly clear, however, that simple univariate analysis is both impractical and does not take full advantage of the rich content of the data and that multivariate analysis methods are increasingly important to effectively mine this data.[1−3] One of the main tasks for which multivariate analysis is used in MSI is to segment different regions of an image for the purpose of diagnosis of diseases or to improve disease understanding and to segment anatomical regions for comparison to histology in order to more fully understand the molecular composition of different anatomical regions.[4,5]

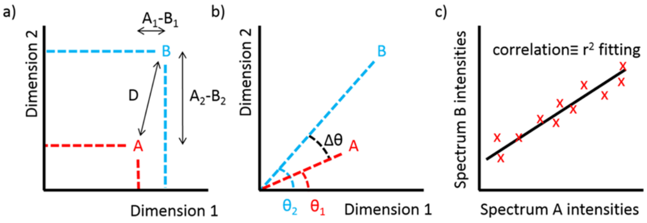
Clustering techniques divide the data set into classes and assign a single class label to each pixel and as such provide a clear categorization of the data. However, the idea of a cluster of data is arbitrary, relying on the notion of "similarity", which can be formulated in many ways. There are many clustering techniques, each of which makes specific assumptions about the data and will therefore categorize a given data set very differently depending on the validity of the assumption.[6−11] There is no *a priori* method for determining which method is appropriate for a given data set. A further and very significant challenge to clustering in MSI is the size of the data itself, both in terms of the number of data points and the dimensionality. A number of different clustering algorithms have been applied to MSI data,[1−3,5,12] each of which makes specific assumptions about the properties of the data and has inherent advantages and disadvantages.[1−3,5,12] Due to its simplicity, relatively low computational requirements,[7] and wide availability in many different languages,[2] *k*-means clustering is one of the most popular algorithms for clustering in MSI.[2,13,14] This can distinguish between anatomies within mouse brain

Received: June 1, 2016
Accepted: September 19, 2016

DOI: 10.1021/acs.analchem.6b02139

tissue[13] and distinguish tumor margins[15] and even intratumor heterogeneity.[4] Given a set of spectra, *k*-means clustering aims to partition *n* spectra into *k* sets to minimize the intracluster sum of distances of each point in the cluster to its cluster center. An illustration of the iterative process of *k*-means clustering is provided in Figure S1 in the Supporting Information.

The distance metric used by the clustering algorithms to compare one spectrum to another (Figure 1) and any normalization strategies applied to the data prior to analysis have a significant effect on the results. In MSI, there can be significant variations in the data that are derived from a number of different experimental sources. For example, variability in sample preparation[16] and laser instability[17] both introduce a source of nonbiological variance within the data. Minimizing these effects by normalization is common but does not and cannot remove all nonbiological variations.[18] Nevertheless, normalization of the data, or pseudonormalization, achieved by the use of the cosine distance, reduces the effects of these variations and thereby improves the clustering results. In the commonly applied TIC normalization, each spectrum is normalized to have unit sum intensity (also referred to as L_1 norm). The cosine similarity is also intensity-independent and therefore also has the potential to reduce the impact of some of these variations on clustering performance (Figure 1b).



$$D = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \qquad D = cos(\Delta\theta) = \frac{A.B}{\|A\|\|B\|} \qquad D = \frac{|A-\bar{A}|.|B-\bar{B}|}{\|A-\bar{A}\|\|B-\bar{B}\|}$$

Figure 1. Visual representations of three of the distance metrics: (a) Euclidean distance, (b) cosine distance, and (c) correlation.

Most applications of *k*-means clustering in MSI have used the Euclidean distance metric,[2,13,14] and where normalization has been used, total ion count (TIC) normalization is the most common.[4,19] Most attempts to evaluate clustering results in MSI have used manual examination or comparison to complementary modalities such as histological analysis.[5] Recently, Oetjen et al. published a series of benchmark 3D data sets with histological information;[20] however, the limited chemical information provided by histology means that segmentations do not always match chemical information provided by MSI.[4]

There are many different methods for quantitatively evaluating the success of clustering, which can be divided into two types: internal and external. Internal evaluation uses the intrinsic properties of the clustering result, usually by comparing the data within each cluster to the data outside of the cluster.[21] Previous attempts to evaluate clustering in MSI have used internal evaluation measures, but these have proven to be inconclusive at best.[22,23] External evaluation on the other hand compares the clustering results to known ground truths such that true and false positives and negative can be computed. Using this information, values such as sensitivity and specificity can be calculated alongside validation measures such as the Rand and Jaccard indices.[24] Since the comparison is to known information, there is no concern of bias toward a given algorithm or distance metric, so it can be used as a method for accurately and reliably comparing and evaluating clustering algorithms or workflows. The main limitation of external evaluation is the need for a ground truth to compare against. Since MSI is generally used as an exploratory tool, usually on biological samples, most data sets will not have a ground truth and thus these external evaluations are usually not possible.[25]

One of the primary assumptions of *k*-means clustering and other algorithms is that the data within clusters is normally distributed. Previously, in other fields, methods have been used to evaluate whether the data within clusters is normally distributed to evaluate the clustering performance[26] or to determine whether to continue to divide clusters further.[27,28] By evaluating the degree of normality within the clusters, when clustering with an algorithm that assumes normality, it is possible to evaluate how well the data fits this assumption and thus how appropriate it is. For univariate data, normality testing is relatively straightforward, and there are a number of tests for normality such as Shapiro−Wilk,[29] Kolmogorov−Smirnov,[30] and Cramer−von Mises[31] tests. This is more challenging in multivariate data since there will be many dimensions, each with different variance and means.[32] However, it is possible to test for multivariate normality using quantile−quantile plots.[33] If the data is multivariate normal, then the Mahalanobis distance will have a χ_p^2 distribution.[34] Therefore, plotting the Mahalanobis distance from each pixel to its relevant distribution versus a χ_p^2 distribution where *p* is the dimensions of the data will give a straight line if the data is multivariate normal.

In this work we show how multivariate normality testing can be used to evaluate the appropriateness of difference distance metrics in *k*-means clustering. We also show how the multivariate normal model can be used as a basis for generating synthetic mass spectrometry imaging data sets, thereby providing samples with a ground truth against which to quantitatively evaluate multivariate analysis methods in MSI as well as other computational analysis methods.

■ MATERIALS AND METHODS

Image Acquisition. Coronal mouse brain was sectioned to 12 μm thickness and thaw-mounted onto glass slides (Superfrost, Thermo Fisher Scientific, Waltham, MA, USA) before being coated with α-cyano-4-hydroxycinnamic acid (CHCA) matrix (5 mg/mL, 80% MeOH 0.1% TFA) using an automated pneumatic sprayer (TM-sprayer, HTX imaging, Chapel Hill, NC, USA). Matrix-assisted laser desorption/ionization (MALDI) images were acquired using a Synapt G2Si (Waters, Manchester, UK), using a pixel size of 45 μm in both the *x* and *y* directions and an *m*/*z* range of 100−1200 Da.

Data Processing and Analysis. Data processing was performed on an Intel Xeon quad core CPU E5-2637 v2 (3.50 GHz) with 64 GB of RAM. All data were converted from a proprietary format to the mzML format using msconvert as part of ProteoWizard[35] software and then into imzML using imzMLConverter.[36] This was then imported into MATLAB (version R2014a and statistics toolbox, The Math-Works, Inc., Natick, MA, USA) using Spectral Analysis.[37] *k*-means clustering was performed using the function *kmeans* from the Matlab Statistics toolbox using the parameters specified in the upcoming experiments and three replicates and random starting clusters. Normality testing was performed on the data within each cluster by plotting the squared Mahalanobis distance from each pixel to the distribution within its cluster against a chi-squared

DOI: 10.1021/acs.analchem.6b02139

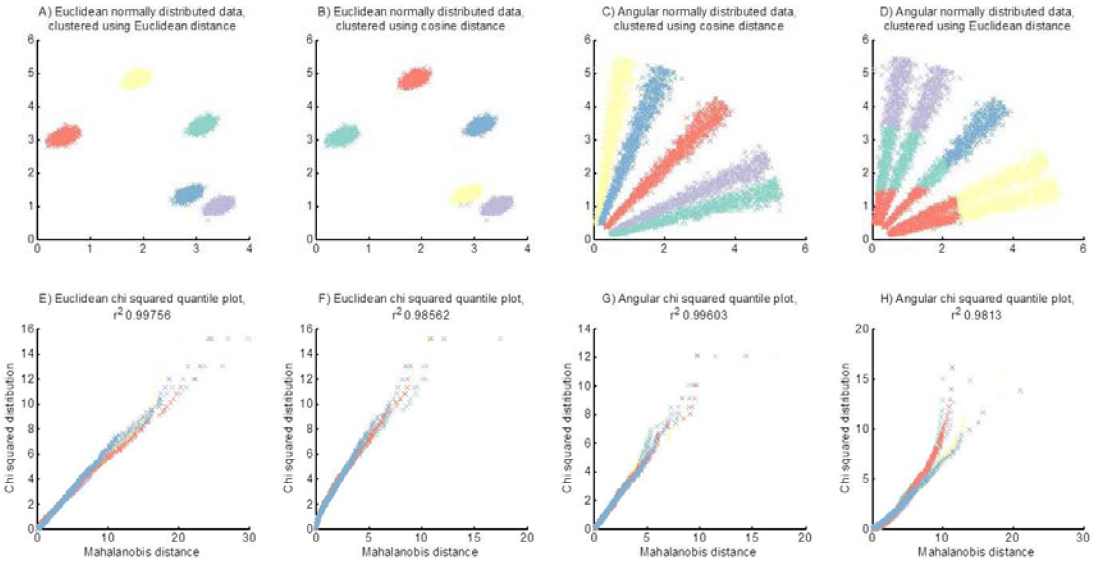


Figure 2. Simulated data of five clusters with normally distributed data (A, B) and angular normally distributed data (C, D), clustered using the Euclidean distance (A, D) and the cosine distance (B, C), with corresponding chi-squared quantile−quantile normality plots below (E−H).

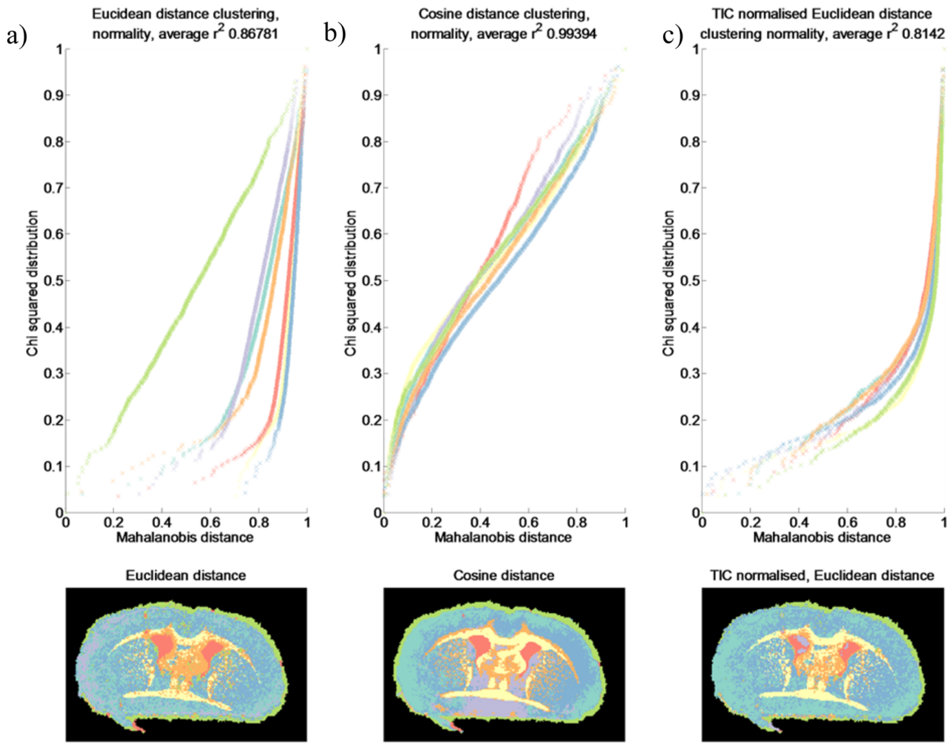


Figure 3. Quantile−quantile plot in (a) Euclidean space (b), angular space, and (c) TIC normalized Euclidean space for the data within each of the 7 clusters of the coronal rat brain image segmented using (a) Euclidean distance, (b) cosine distance, and (c) Euclidean distance with TIC normalization.

distribution with a number of degrees of freedom equal to the dimensions of the data.[33] The Mahalanobis distance for the data within each cluster was calculated by first performing PCA, removing components with zero variance and scaling such that each component has a standard deviation of 1, and then calculating the squared Euclidean distance of each pixel to the mean of its assigned cluster.[38] For data clustered using the cosine distance metric, data were first converted into a polar coordinate system, comprising a distance from the origin r and a series of angles from the origin θ_{n-1} relative to each of the coordinate axes where n is the dimensionality of the data.[39] The angles from the coordinate axes were then used to determine normality of the angular distribution. For creating the plots, the Mahalanobis distance and chi-squared values were all rescaled to between 0 and 1 in order to plot them all on a common axis.

Synthetic data were generated using the following workflow:

1. Convert the reference data to polar coordinates.
2. Test for normality of the reference data in polar coordinates using chi-squared quantile−quantile plotting.
3. If the reference data is multivariate normal, then calculate the means and covariances of the reference data in polar coordinates.

DOI: 10.1021/acs.analchem.6b02139

4. Generate a set of synthetic multivariate normally distributed data with the mean and covariance of the reference data using the *mvnrnd* function in MATLAB.
5. Convert the synthetic multivariate normal data back to Cartesian coordinates.
6. Populate a spatial mask with the synthetic data.

■ RESULTS AND DISCUSSION

Two synthetic data sets were generated to simulate data that is normally distributed in Euclidean and polar coordinates, respectively. Clustering was performed using *k*-means with both the Euclidean and cosine distances. Normality testing via quantile−quantile plots revealed that both synthetic data sets clustered well under the cosine distance, whereas the data distributed normally in polar coordinates did not cluster well when the Euclidean distance was used (Figure 2).

k-means clustering was then performed on an MSI image from coronal mouse brain with $k = 2−10$; 7 clusters were then chosen based on visual assessment of the resulting images, along with comparison to the Allen brain atlas. When applied to an MSI image of a biological system (coronal mouse brain), the chi-squared quantile plots show that the data within clusters obtained using the cosine distance have a higher r^2 value than the data within the clusters using the Euclidean distance (average r^2 0.99 compared to 0.87; Figure 3A,B). This means that the data in the clusters formed using the cosine distance are closer to normally distributed than the Euclidean distance. This indicates that the cosine distance is the more appropriate distance metric for cluster with on this data set based on the multivariate normal assumption of the *k*-means algorithm. The inappropriateness of *k*-means with the Euclidean distance in this case mirrors the visually poor results obtained with respect to the anatomical features expected from coronal mouse brain as seen in the Allen brain atlas (Figure 3).[40] In comparison, the cosine distance gives visually clearer results, and the distribution of points within clusters are more normally distributed in the appropriate space. We note that use of the common TIC normalization *decreases* the normality of the data and does not produce visually clearer segmentation images (Figure 3c). The reason for this is that TIC normalization rescales all data points such that they lie on the surface of a hyperdiamond (lines of constant L_1 norm). Thus, they are certainly not normally distributed as one dimension is condensed. They might be normally distributed if you consider only positions on the hypersurface. Results obtained from additional data sets (sagittal rat brain and mouse lung tissue) produce similar results with respect to comparison of normality to visual appearance of clustering results and are provided in the Supporting Information (Figures S2 and S3). It is worth noting that the values produced from the r^2 fitting cannot easily be directly interpreted, as it will be dependent on the number of data points and the dimensionality of the data. Therefore, it is recommended as a means to compare results, and caution should be taken when inferring additional information from them.

It is also worth noting that the shape of the quantile−quantile plots are not completely linear, a feature that can arise from a number of different sources. For example, the presence of a few outliers will skew the distribution toward a sigmoidal shape, as is observed when the cosine distance is used (Figures S4 and S5). This is caused by the outliers skewing the mean of the data and thus altering the Mahalanobis distance for every point. While this effect is minimized through the variance scaling process, some effect can still be observed. Alternately, a circular distribution of data with a core of normal data within produces a similar shaped, apparent bilinear plot to those observed when using the Euclidean distance (Figure S6). This is not indicative of two normally distributed sets of data, which produce a different shaped plot (Figure S8). For further examples of how other distributions of data will affect these plots, see Figures S4−S12. However, we note that caution is required when generalizing from these plots from two-dimensional data into the higher dimensional space in which MSI data sits.

While distance metric determination is a crucial factor in any clustering algorithm, there are still many other parameters that must also be selected, such as the number of clusters and the method for centroid initiation. In addition to this, there are many other clustering approaches, such as density-based clustering, that do not assume multivariate normality in the data. Therefore, a method to generate data sets with a ground truth is required to assess the suitability of these approaches and to permit a comparison of different clustering approaches. Data simulated from first-principles is one approach that is used to achieve this in other fields. However, while some aspects of image formation and noise in MSI are well-understood, there are still a large number of unknowns in aspects such as sample preparation and ionization.[41,42] One approach is to take existing peak lists and to then simulate the known variables and apply these to this peak list.[43] However, a robust method is needed to generate peak lists that are well-controlled but still representative of the variance expected from biological samples. A new biological sample could be analyzed each time a new set of spectra are required, but using new animal or human tissue each time a different number of regions or pixels is required is neither practical or ethical. In other areas, such as financial prediction and geological analysis, statistical modeling is used to convert discrete data into a continuous function, thereby allowing resampling to generate the desired number of data points. Statistical modeling assumes that data from a population are derived from a known probability distribution function. Provided that the model adequately describes the data, the underlying distribution can then be resampled to give a new synthetic data set with any desired number of data points. This new synthetic data set will have the same distribution as the original reference data set from which the model was derived. For large and high dimensional data, model generation and parameter estimation can be challenging; however, the multivariate normal model parameters can be easily estimated even for very large data sets.[44] As previously demonstrated, clustered MSI data closely approximates to a multivariate normal distribution when the data is converted to polar coordinates. This means that the multivariate normal distribution can be used as the basis for statistical modeling for MSI data. The small deviations from normal are most probably due to a few outlier pixels within the data rather than a deviation from normal within the majority of the data itself. This is demonstrated by the similarity of the sigmoidal nature of the plot (Figure 3B) to that generated using simulated outliers (Figures S4 and S5). This suggests that synthetic spectra generated in this way are representative of those observed in the real data and thereby serve as a basis to introduce and explore additional experimental or instrumental variabilities in a controlled way.

In order to perform statistical modeling of MSI data, a series of seven anatomical features from an MSI image of the previously shown mouse brain were used as a reference data set (Figure 4). These regions were generated based on the analysis of selected ion images and PCA scores, in comparison to a high-resolution optical image (Figure S13) and the Allen brain atlas.[40] These data

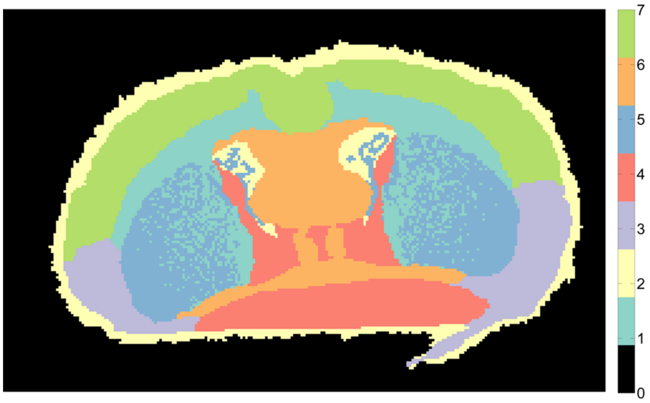


Figure 4. Seven anatomical regions used as reference data for statistical modeling, segmented based on a combination of comparison of a high-resolution optical image (Figure S13) with selected ion images and PCA scores images along with comparison to the Allen brain atlas.

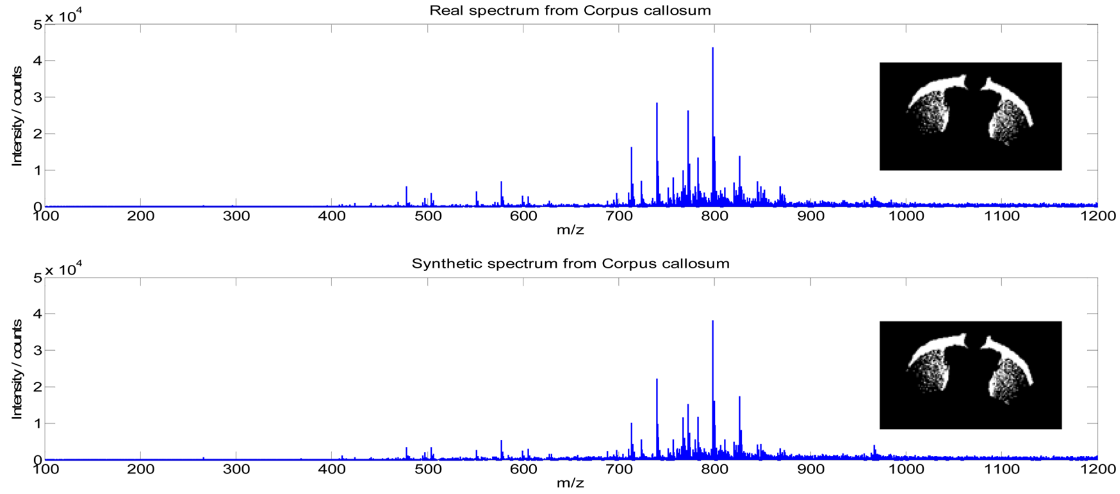
sets were then tested for normality in polar and Euclidean spaces using the chi-squared quantile plots shown previously and showed a high degree of normality throughout polar space but not Euclidean (Table 1). This suggests that the multivariate normal model can be used to summarize the properties of this data. From this model, a new synthetic data set was generated by resampling from the distribution with the same number of pixels as the original reference data. The synthetic spectra from a number of the different anatomical regions were then visually compared to the original reference spectra (Figure 5). The synthetic and real spectra show a high degree of spectral similarity, and expected features such as isotope ratios and fragments are preserved, thus ensuring the realism of the synthetic data. Some differences in the spectra are observed, since the synthetic spectra are sampled from a distribution and will therefore contain the same underlying variance as the reference data. This is important since biological samples vary, so in order to be realistic, the synthetic data must incorporate this variance.

Table 1. Results of r^2 Fitting of the Chi-Squared Quantile−Quantile Plots on the Seven Anatomical Regions in Figure 4 Showing That the Data Is Closer to Normal in Polar Space than in Euclidean

ID	region	polar normality (r^2)	Euclidean normality (r^2)
1	corpus callosum	0.983	0.822
2	outer boundary	0.983	0.904
3	olfactory areas	0.993	0.889
4	brain stem	0.988	0.854
5	caudoputamen	0.994	0.948
6	lateral septal complex	0.984	0.916
7	isocortex	0.990	0.680

Visual comparison of the spectra is insufficient to evaluate data sets that will be analyzed by multivariate methods. Therefore, in order to evaluate how closely the synthetic data matches real data, a new data set, comprising both synthetic and real spectra, was generated. Principal component analysis (PCA) was then performed on this combined data set to determine if the statistical modeling process introduced any additional observable variance. No principal component scores were found to separate the synthetic from real data (Figures 6 and S14). This means that even when all mass channels are considered the difference between the synthetic and real data is smaller than that between different anatomies or the spectral noise within the data and supports the suggestion that the differences from normal are likely to be outlier pixels. As such the statistical modeling of appropriately segmented MSI data using a multivariate normal distribution can generate realistic spectra in order to create new data sets with known ground truth for external evaluation of clustering in mass spectrometry imaging.

Large synthetic data sets can also be generated rapidly using this approach, by simply taking more samples from the multivariate normal distribution. To demonstrate this, a data set containing 9 times the number of pixels of the original reference data was generated (187 452 pixels from 20 825 in the original). This represents the size of data from an area 3 times the size in each dimension or if the image had been acquired with 15 μm rather than 45 μm pixels. These new data were generated in approximately 5 min, but it would have required around 36 h to acquire the same number of pixels experimentally. PCA performed on a combined data set containing the new larger data set and the original reference data still shows no separation between the synthetic and real data, demonstrating that this approach scales to large data sets without any statistically


Real spectrum from Corpus callosum

Synthetic spectrum from Corpus callosum

Figure 5. (Top) Real spectrum from the corpus callosum region of the reference data. (Bottom) Synthetic data sampled from the multivariate normal distribution of the corpus callosum region. A high degree of spectral similarity is observed between the spectra.

DOI: 10.1021/acs.analchem.6b02139

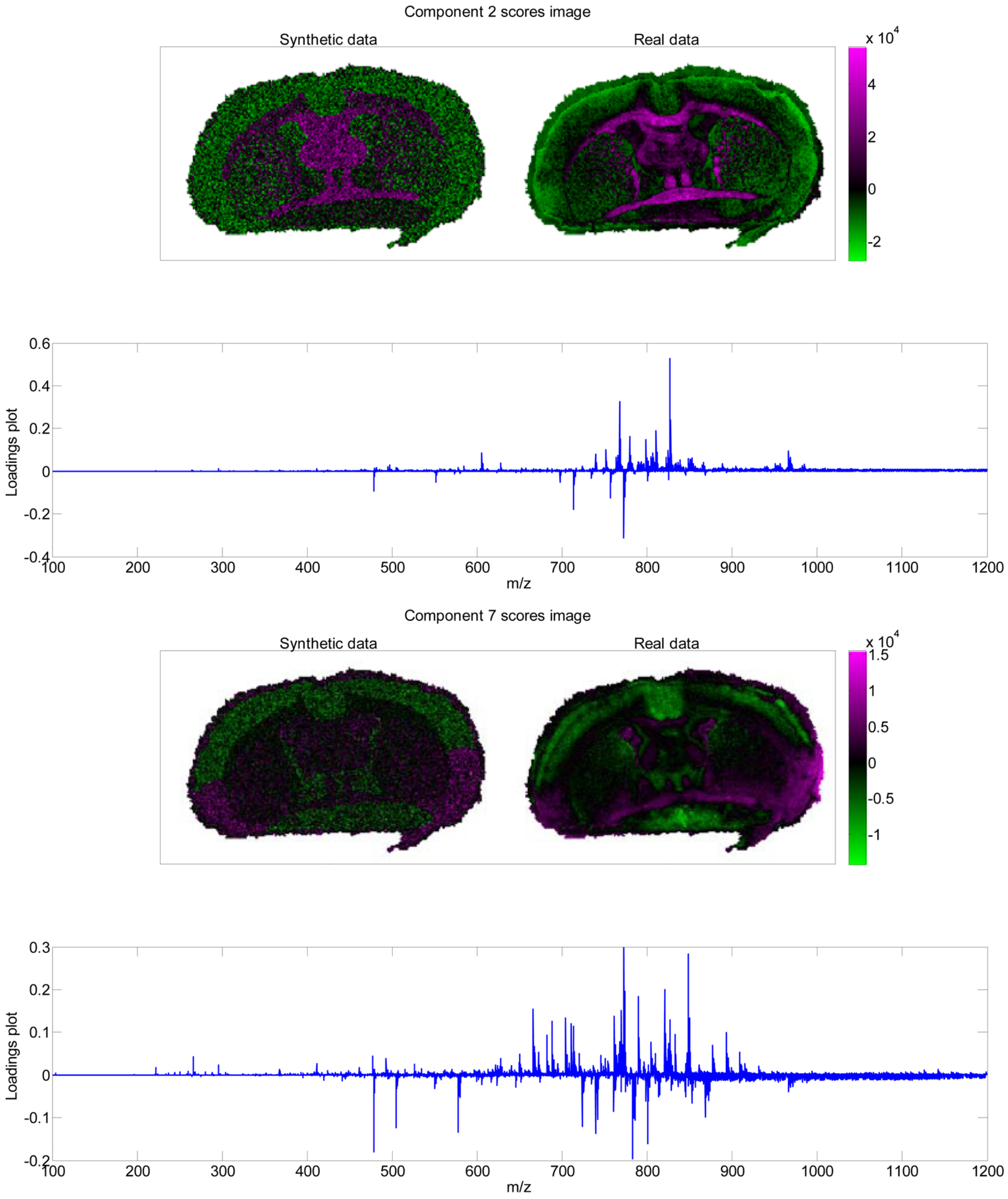


Figure 6. Results of PCA on the combined data set containing both real and synthetic data, with scores images on top and projection loadings plots on the bottom. No principal component was found to separate the real from the synthetic data, indicating that any variance in the data is from the inherent biological and experimental variance in the reference data rather than introduced by the statistical modeling.

detectable changes occurring in the data (Figure S15). While in both of these cases the full seven regions were used to generate synthetic data, an image containing any desired number of regions can be generated using this approach, provided there is a suitable set of reference data. This means that the performance of different clustering algorithms or multivariate analysis methods can be evaluated with respect to the size and complexity of the data in terms of expected features. In addition, no new tissue sections are required, allowing the potential to minimize animal usage in computational studies in MSI. We note that the synthetic images appear more speckled than the reference data. This is because no spatial smoothing is applied and neighboring pixels are statistically independent when populating the spatial masks with spectra. This could potentially be overcome by also maximizing the similarity of neighboring pixel, but for clustering evaluation this is unnecessary.

■ CONCLUSIONS

Robust evaluation of clustering in MSI allows us to understand its limitations and what can be deduced from its results. In the case of *k*-means clustering and other algorithms that assume normality of the data (such as agglomerative hierarchical clustering with Ward's linkage), we have shown that, in the absence of ground truth data, evaluation of multivariate normality in the intracluster distributions is an internal test that can be used on MSI data to determine, postanalysis, whether clustering should be performed using the cosine or Euclidean distance. Where possible, external evaluation methods should be used when comparing novel algorithms or parameters, using a ground truth that is representative of samples of interest. We have demonstrated that synthetic data generated by statistical modeling is a suitable means to achieve this. In addition, this approach allows large data sets to be generated rapidly, allowing evaluation and comparison of both existing and new methods as the data increases in size.

■ ASSOCIATED CONTENT

S Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b02139.

Further detail about the *k*-means algorithm and examples of synthetic data of varying normalities (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: josephine.bunch@npl.co.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was funded by NPL strategic research programmes 116301 and 117194. The authors thank Dr. Rory Steven for his assistance with MALDI MS imaging experiments and Professor Helen Cooper for useful discussions. A.D. gratefully acknowledges financial support from the EPSRC through a studentship from the PSIBS Doctoral Training Centre (EP/F50053X/1) in collaboration with the National Physical Laboratory and AstraZeneca. Data supporting this research is openly available from the University of Birmingham data archive at http://findit.bham.ac.uk/.

■ REFERENCES

(1) Muir, E.; Ndiour, I.; Le Goasduff, N.; Moffitt, R. A.; Liu, Y.; Sullards, M. C.; Merrill, A. H.; Chen, Y.; Wang, M. D. *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering*; IEEE, 2007; pp 472−479.

(2) McCombie, G.; Staab, D.; Stoeckli, M.; Knochenmuss, R. *Anal. Chem.* **2005**, *77*, 6118−6124.

(3) Fonville, J. M.; Carter, C. L.; Pizarro, L.; Steven, R. T.; Palmer, A. D.; Griffiths, R. L.; Lalor, P. F.; Lindon, J. C.; Nicholson, J. K.; Holmes, E.; Bunch, J. *Anal. Chem.* **2013**, *85*, 1415−1423.

(4) Jones, E. A.; van Remoortere, A.; van Zeijl, R. J.; Hogendoorn, P. C.; Bovée, J. V.; Deelder, A. M.; McDonnell, L. A. *PLoS One* **2011**, *6*, e24913.

(5) Deininger, S. r.-O.; Ebert, M. P.; Fütterer, A.; Gerhard, M.; Röcken, C. *J. Proteome Res.* **2008**, *7*, 5230−5236.

(6) Estivill-Castro, V. *SIGKDD Explor.* **2002**, *4*, 65−75.

(7) Hartigan, J. A.; Wong, M. A. *Appl. Stat.* **1979**, *28*, 100−108.

(8) Jain, A. K. *Pattern Recogn. Lett.* **2010**, *31*, 651−666.

(9) Birant, D.; Kut, A. *Data Knowl. Eng.* **2007**, *60*, 208−221.

(10) Fu, L.; Medico, E. *BMC Bioinf.* **2007**, *8*, 3.

(11) Choong, M. Y.; Kow, W. Y.; Chin, Y. K.; Angeline, L.; Teo, K. T. K. *Proceedings of the IEEE International Conference on Control System, Computing and Engineering*; IEEE, 2012; pp 430−435.

(12) Trede, D.; Schiffler, S.; Becker, M.; Wirtz, S.; Steinhorst, K.; Strehlow, J.; Aichler, M.; Kobarg, J. H.; Oetjen, J.; Dyatlov, A.; Heldmann, S.; Walch, A.; Thiele, H.; Maass, P.; Alexandrov, T. *Anal. Chem.* **2012**, *84*, 6079−6087.

(13) Race, A. M.; Steven, R. T.; Palmer, A. D.; Styles, I. B.; Bunch, J. *Anal. Chem.* **2013**, *85*, 3071−3078.

(14) Alexandrov, T.; Becker, M.; Deininger, S. O.; Ernst, G.; Wehder, L.; Grasmair, M.; von Eggeling, F.; Thiele, H.; Maass, P. *J. Proteome Res.* **2010**, *9*, 6535−6546.

(15) Alexandrov, T.; Becker, M.; Guntinas-Lichius, O.; Ernst, G.; von Eggeling, F. *J. Cancer Res. Clin. Oncol.* **2013**, *139*, 85−95.

(16) Goodwin, R. J. *J. Proteomics* **2012**, *75*, 4893−4911.

(17) Steven, R. T.; Dexter, A.; Bunch, J. *Methods* **2016**, *104*, 101−110.

(18) Deininger, S.-O.; Cornett, D. S.; Paape, R.; Becker, M.; Pineau, C.; Rauser, S.; Walch, A.; Wolski, E. *Anal. Bioanal. Chem.* **2011**, *401*, 167−181.

(19) Abdelmoula, W. M.; Carreira, R. J.; Shyti, R.; Balluff, B.; van Zeijl, R. J.; Tolner, E. A.; Lelieveldt, B. F.; van den Maagdenberg, A. M.; McDonnell, L. A.; Dijkstra, J. *Anal. Chem.* **2014**, *86*, 3947−3954.

(20) Oetjen, J.; Veselkov, K.; Watrous, J.; McKenzie, J. S.; Becker, M.; Hauberg-Lotte, L.; Kobarg, J. H.; Strittmatter, N.; Mróz, A. K.; Hoffmann, F. *GigaScience* **2015**, *4*, 1−8.

(21) Liu, Y.; Li, Z.; Xiong, H.; Gao, X.; Wu, J. *Proceedings of the IEEE 10th International Conference on Data Mining*; IEEE, 2010; pp 911−916.

(22) Van de Plas, R.; Ojeda, F.; Dewil, M.; Van Den Bosch, L.; De Moor, B.; Waelkens, E. *Bioinformatics* **2006**, 1−10.

(23) Sarkari, S.; Kaddi, C. D.; Bennett, R. V.; Fernandez, F. M.; Wang, M. D. *Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*; IEEE, 2014; pp 4771−4774.

(24) Rand, W. M. *J. Am. Stat. Assoc.* **1971**, *66*, 846−850.

(25) Garden, R. W.; Sweedler, J. V. *Anal. Chem.* **2000**, *72*, 30−36.

(26) Mao, J.; Jain, A. K. *IEEE Trans. Neural Netw.* **1996**, *7*, 16−29.

(27) Hamerly, G.; Elkan, C. *Adv. Neural Inf. Process. Syst.* **2004**, *16*, 281.

(28) Steinley, D. *Br. J. Math. Stat. Psychol.* **2006**, *59*, 1−34.

(29) Shapiro, S. S.; Wilk, M. B. *Biometrika* **1965**, *52*, 591−611.

(30) Lilliefors, H. W. *J. Am. Stat. Assoc.* **1967**, *62*, 399−402.

(31) Darling, D. A. *Ann. Math. Stat.* **1957**, *28*, 823−838.

(32) Goeman, J. J.; Van De Geer, S. A.; Van Houwelingen, H. C. *J. R. Stat. Soc. Series B Stat. Methodol.* **2006**, *68*, 477−493.

(33) Burdenski, T. K., Jr *Multiple Linear Regression Viewpoints* **2000**, *2*, 15−28.

(34) Healy, M. *Appl. Stat.* **1968**, *17*, 157−161.

(35) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. *Bioinformatics* **2008**, *24*, 2534−2536.

(36) Race, A. M.; Styles, I. B.; Bunch, J. *J. Proteomics* **2012**, *75*, 5111−5112.

(37) Race, A. M.; Palmer, A. D.; Dexter, A.; Steven, R. T.; Styles, I. B.; Bunch, J. *Anal. Chem.* **2016**, *88* (19), 9451−9458.

(38) Mahalanobis, P. C. *Proc. Natl. Inst. Sci. (Calcutta)* **1936**, *2*, 49−55.

(39) Kendall, M. G. *A Course in the Geometry of n Dimensions*; Courier Corporation, 2004.

(40) Lein, E. S.; Hawrylycz, M. J.; Ao, N.; Ayres, M.; Bensinger, A.; Bernard, A.; Boe, A. F.; Boguski, M. S.; Brockway, K. S.; Byrnes, E. J. *Nature* **2007**, *445*, 168−176.

(41) Ipsen, A. *Anal. Chem.* **2015**, *87*, 1726−1734.

(42) Du, P.; Stolovitzky, G.; Horvatovich, P.; Bischoff, R.; Lim, J.; Suits, F. *Bioinformatics* **2008**, *24*, 1070−1077.

(43) Palmer, A. D. *Information processing for mass spectrometry imaging.* Ph.D. Thesis, University of Birmingham, 2014.

(44) Xu, J. J. *Retrospective Theses and Dissertations* **1996**, 3120.