Supporting Information

Chemometric analysis of MALDI mass spectrometric images of three-dimensional cell culture systems

The material presented in the supporting information below describes an introduction to the theory of several chemometric methods. Generally, the material is presented on a “need to know” basis to those without significant knowledge of the field and requires only a basic knowledge of introductory statistics. Therefore, occasionally details/specifcs of the techniques are omitted for clarity and oversimplifications are made to improve flow. There are several introductory texts cited within this material that can be used by readers for further information and study.
Principal Component Analysis

Data sets that record one independent variable are termed univariate. Examples of univariate data sets include obtaining a single mass spectrum at one pixel within a MALDI image (the independent variable is $m/z$) or monitoring the intensity at one specific $m/z$ across all pixels within an IMS experiment (the independent variable is location). Multivariate data sets contain more than one independent variable per measurement element. Examples of multivariate data include collecting entire mass spectra at discrete points in time within separation methods (i.e. scan mode HPLC-MS where time and $m/z$ are simultaneously two independent variables) or collecting entire mass spectra across several pixels within an IMS experiment (where location and $m/z$ are simultaneously two independent variables). Principal component analysis (PCA) is an exploratory data analysis tool used to identify statistically significant sources of variation within multivariate data sets. Stated another way, PCA can be used to identify trends within data sets consisting of multiple independent spectra. A brief overview will be presented here as a guide for the data being presented, but readers are referred elsewhere for a full review\textsuperscript{1,2}.

The goal of PCA is to mathematically simplify the representation of a data set and can be best presented using an example. Consider a MALDI IMS data set compromised of individual mass spectra recorded for a set of pixels within an image. PCA identifies spectral features termed \textit{principal components} (also called factors, eigenvectors, or loadings) that are common within pixels of a data set. Some of these principal components describe relevant information such as a peak at a particular $m/z$ value while others only describe noise. Principal components are considered independent variables in PCA much the same way $m/z$ values are considered independent variables in a recorded mass spectrum. In addition to loadings, PCA also identifies \textit{scores} which are a set of scalar numbers that describe how important each spectral trend is to each data element. In terms of MALDI IMS data, scores represent how important each spectral trend is to each pixel. A large positive score value indicates that a
spectral trend is strongly positively correlated with a particular pixel while a large negative score value indicates that a spectral trend is strongly negatively correlated with a particular pixel. The entirety of a measured mass spectrum for a particular pixel can be reconstituted by summing the multiplicative products of each principal component with a corresponding score. A graphical view of this process is shown in Supplemental Figure S1.

Supporting Information Figure S1. Example of PCA. Consider a spheroid slice analyzed by MALDI IMS. A mass spectrum taken from one pixel of a MALDI IMS data set (black, left) can be represented by a set of spectral features (colored, right) after performing PCA. PCA identifies several spectral features termed principal components (PCs) associated within all pixels of this particular data set (red, blue, green, and magenta graphs). A set of scores (numbers inset within each graph) describe how important each of these principal components are to the mass spectrum of this particular pixel. The original data can be reconstituted by multiplying each principal component loading by its score and summing the result.

Note that principal components must never be considered individual, pure analyte spectra that are somehow summed to give an overall recorded spectra; principal components are abstract representations that highlight trends present within the data and are not always directly interpretable\textsuperscript{3, 4}. In addition, the breakdown in Figure S1 shows one main assumption made by PCA: the overall measured response (e.g. intensity, counts, etc.) is considered a linear sum of all responses from individual contributors. In terms of MALDI IMS data, this means that summing
the mass spectra from individual analytes would generate the actual measured mass spectrum at a particular pixel. PCA would not be a suitable technique to use if the measured signal originated from the multiplicative product of individual spectra.

The PCA algorithm successively calculates independent principal components in order of importance, with the first several principal components describing the prominent spectral features within the data and the latter principal components describing spectral minutia or noise. The principal components that describe relevant spectral information are retained and used for further data processing while those that describe noise are discarded with negligible impact as shown in Supplemental Figure S2. As such, another assumption of PCA is that the meaningful information has a significantly higher measured response apart from noise. In other words, meaningful PCA requires data with at least a moderately high signal-to-noise ratio.

Supporting Information Figure S2. Noise Removal by PCA. The original spectrum from Supporting Figure S1 (black trace) can be reconstructed nearly identically using only the first several principal components that describe relevant information (red trace). The latter principal components describe only noise (purple trace) and can be discarded.

There are a variety of approaches for separating the principal components that describe relevant information from those that describe noise\(^2,6\). The Scree graph was used here\(^2,6\). In simple terms, this approach plots the relative amount of information (termed variance) described by each principal component. Principal components that are relevant should theoretically
contain more variance than those that describe noise. The variance is calculated for each principal component, presented on a logarithmic scale, and an “elbow” in the graph is taken to be the maximum number of relevant principal components to retain.

A score plot is a graphical display often used with PCA. In this type of plot the magnitude of scores for each pixel along specified principal components are plotted (e.g. PC2 versus PC1). For MALDI IMS data, each point in this score plot represents the spectral content described by the plotted principal components for a particular pixel. Stated another way, since each principal component is a multivariate independent variable, the spectral content of a particular pixel is represented as a single point in an abstract multivariate coordinate system. Theoretically, the points associated with pixels that have similar mass spectral content should cluster together while points associated with pixels that have dissimilar spectral content should be segregated. This idea of transforming entire mass spectra into single points whose spatial distribution in a PCA score plot is associated with spectral similarity serves as the foundation for the clustering analysis (see below). Finally, much the same way the single points represent spectra within a PCA score plot, any line drawn within a PCA score plot also represents an entire spectrum of information (since single points define a line). In this case, the axis lines representing principal components are the spectral trends identified by PCA. In sum, a PCA score plot of MALDI IMS data graphically plots the mass spectra of individual pixels according to how important/evident certain spectral trends (principal components) are exhibited within each of the pixels.

Mean-centering and variance scaling are two general data pre-processing steps that are often performed prior to PCA. Mean-centering is the process by which an average data spectrum is calculated and subtracted from each spectrum of the data set. Graphically, this process will center the data points within the PCA score plot about the origin. Without mean-centering, the first principal component calculated will simply be the average of the data set and will not describe variation in spectral content7.
Mathematical scaling procedures are often used to ensure that all variables contribute equally during the PCA calculation. Intensities vary dramatically in mass spectra as a function of $m/z$. Higher intensity peaks typically have a numerically wider range of intensities across pixels as compared to lower intensity peaks. Stated another way, the absolute standard deviation of peak amplitudes across all pixels for a common high intensity peak would be larger than the standard deviation of peak amplitudes across all pixels for a common low intensity peak. The $m/z$ values associated with peaks that exhibit wide ranges in intensity describe more of the total variation of the entire MALDI IMS data set while $m/z$ values associated with peaks that exhibit smaller ranges in intensity describe less of the total variation. Since PCA is a technique that is used to identify sources of spectral variation in successive order of importance, performing PCA without scaling would cause high intensity peaks to dominate the calculated spectral trends/principal components at the expense of significant, yet lower intensity peaks.

A simple scaling method can be performed by dividing the intensity at a specified $m/z$ for a specific pixel by the variance (square of standard deviation) of its intensity values across all pixels. Unfortunately, scaling often enhances noise. By treating all $m/z$ values as equally important, $m/z$ values that are associated solely with noise are given equal weight to the $m/z$ values that describe relevant spectral features. Thus, variance scaling can sometimes degrade the quality of the data set being analyzed.

**Cluster Analysis**

The goal of cluster analysis is to identify and group similar observations present within a data set. Clustering is termed an *unsupervised* data analysis technique because the methodology is used without any *a priori* knowledge or assumptions of the data set (e.g. the number of groups that exist, the proper groupings of elements within a data set, the structure/distribution of the data, etc.). Two clustering methods are discussed below: hierarchical clustering analysis (HCA) and $k$-means clustering.
In HCA, each point is initially treated as an individual cluster and inter-point distances are calculated for all data points. While there are several inter-point distance measures, the most commonly used is Euclidean distance, or the square root of the sum of squared coordinate point differences (i.e. the Pythagorean Theorem for a two-dimensional coordinate system and similar extensions for higher-dimensional coordinate systems). After inter-point distances are calculated, the two nearest points are merged together and treated as a single cluster entity, the inter-point distances across the data set are again calculated, and again, observations are merged. This process continues in an iterative fashion whereby either two separate points are joined, a point is joined with a cluster, or two clusters are joined until all data elements are interconnected into a nested hierarchy.

There are several different methods in HCA that can be used to calculate the distance or linkage between two different clusters. Depending on the spatial distribution of the data, sometimes these different linkage measures can generate alternative (although not necessarily incorrect) groupings of the points within the data set. Two common linkage measures in HCA include nearest-neighbor (termed single linkage) and furthest-neighbor (termed complete linkage). In single linkage, the distance between two neighboring clusters is taken to be the shortest distance between two points from the two distinct clusters. Single linkage suffers from two disadvantages. First, grouping by the nearest neighbor ignores the underlying distribution of the data set. Second, a phenomenon termed chaining can occur whereby grouping errors arise because some data points can be close enough to one another to allow grouping, yet can still be far away from most of the elements of the group; graphically, HCA with nearest-neighbor linkage can thus include points within a chain-link type distribution pattern away from a center of mass. In complete linkage (also known as furthest-neighbor) the distance between two neighboring clusters is taken to be the farthest distance between two points from the two separate clusters. Analyses performed with complete linkage tends to bias towards creating
clusters with a spherical distribution. Linkages based on average distances between points in different clusters and cluster centroids can also be used.

Regardless of the linkage strategy, once all data observations are grouped into a nested hierarchy, the user can estimate the likely number of separate groups that exist based on a distance threshold where sets of data elements must be a minimum distance apart to be considered separate clusters. An excellent visual representation of this process is shown in Figures 4.10, 4.11, and 4.15 of Beebe.

$k$-Means clustering is an iterative method that divides data elements into a pre-defined number of groups (N) by minimizing the sum-of-squares spread of the cluster about its center. By minimizing the spread about the cluster center, $k$-means clustering assumes spherical cluster distributions. This process is shown graphically in Supplemental Figure S3. Within the data set to be clustered, N points are first randomly chosen as cluster centers (also called “seeds”) and all other data elements are assigned into the group with the nearest cluster center. Next, the center of mass of the grouped elements (also called a centroid) is calculated and the process of assigning elements to the cluster with the nearest center is repeated. These two steps of calculating cluster centroids and assigning groupings are alternated iteratively until the result no longer changes (convergence is reached). Similar to HCA, there are several methods for calculating inter-point distances (which determine data assignments into clusters and cluster centroid locations), with Euclidean distance being the most common.
**Supporting Information Figure S3. k-Means Clustering.** A set of data was generated that contained two clusters of points with random noise added (A). The algorithm randomly chooses two cluster seeds to start the process (circled red and blue in A). All other points in the data set are assigned to the nearest cluster (B). Using these associations, new cluster centers are calculated, represented by the two Xs in B. The process of assigning points to clusters is repeated again as shown in C. Note how the cluster assignments changed. The calculation of cluster centers and point assignments is repeated until convergence is reached (D).

*k*-Means clustering suffers from two main disadvantages. First, the assignment of data elements to clusters is highly dependent on the starting location of the cluster centers. To address this issue, the entire *k*-means clustering process is repeated multiple times on the same data set using different starting seeds to ensure optimal clustering. Second, and more important, the number of clusters within the data set must be known and assigned before the clustering begins.
There are several methods that can be used to estimate the proper number of clusters without any \textit{a priori} knowledge of a data set. In the gap-statistic method, the spread (also termed dispersion) of points within groups of a clustered data set is compared against the spread of points within a data set containing random numbers. As an example, assume a MALDI IMS data set ideally contains four separate groups of pixels that exhibit four unique mass spectral profiles. If fewer than four groups are used to cluster this “structured” data set, the dispersion of points within clusters will be large because not enough clusters exist to properly group the data points. In essence, the span of the too few clusters must increase to “reach out” farther to encompass all the data points. Conversely, if more than four groups are used, intact clusters are usually just split into smaller pieces, exhibiting a very small decrease to the within-cluster spread. For an unstructured (also called null) data set consisting of only random numbers that have no clusters, dispersion should gradually decrease at a constant rate (more or less) as the number of clusters used to group the data elements increase. The “gap” between the dispersion of a data set consisting of random numbers and the data set being tested is then used to estimate of the proper number of clusters for the test set.

\textbf{Principal Component Analysis – Linear Discriminant Analysis}

The field of discriminant analysis is traditionally used to classify unknown data elements into pre-defined groups based on data similarity. For example, assume MALDI IMS is used to monitor protein expression of four known cell types spatially localized within four distinct regions (e.g. north, south, east, and west) on the surface of an imaging plate. If each of these four distinct cell types contains a unique protein expression pattern, then (theoretically) four unique, characteristic mass spectral profiles should be generated. If these same four cell types are then randomly distributed across another imaging plate, discriminant analysis is an approach that can be used to associate each pixel of this new image with a specific cell type based on the similarity of their mass spectra with those unique, characteristic mass spectra previously
recorded. Discriminant analysis is considered a *supervised* technique because information must be known by the user before classification can begin: the proper number of groups that exist within a data set and their corresponding group labels/identifiers (e.g. proliferating cells, necrotic cells, etc.).

Linear discriminant analysis (LDA) can best be described using an example. Consider the two spatially-segregated groups of points identified by $k$-means clustering shown in Supplemental Figure S3. The LDA algorithm calculates the equation of a line that best distinguishes between the two groups of data; this line (also called a classifier, linear boundary, or linear discriminant) separates the graph into two distinct regions as shown in Supplemental Figure S4. The line calculated by LDA maximizes intergroup variance and minimizes intragroup variance, all with the lowest possible misclassification error. When any future observation of unknown association is plotted in this coordinate system, it would be classified as either one group or another according to its location on either side of the linear boundary. LDA assumes that each of the data groups has a Gaussian distribution and the groups have the same covariance (a measure of the strength of interaction among the plotted variables).

**Supporting Information Figure S4. Calculation of a Linear Boundary by LDA.** The data shown here is taken from Supporting Information Figure S3D. LDA calculates the equation of a boundary (black line) that segregates the coordinate system into two regions (orange for cluster 1 and green for cluster 2). Any future observations would be categorized into a specific group depending on its location on either side of the line.
As opposed to just a “line”, LDA more broadly refers to a linear combination of input variables (with exponents of one as opposed to a quadratic or cubic discriminant which would have exponents of two or three, respectively). In two dimensions, LDA generates a line because a line is the linear combination of the two independent input variables X and Y:

\[ AX + BY = C \]  

(Eq. S1)

where A, B, and C are scalars (A and B are scalar coefficients) used to determine the direction of the line. This linear combination is not limited to only two measured variables. A plane would be created by LDA in a three-dimensional coordinate system because a plane is the linear combination of three input variables X, Y, and Z:

\[ AX + BY + CZ = D \]  

(Eq. S2)

While impossible to graph, LDA would continue to generate a linear-based boundary for data with larger numbers of input variables:

\[ AX + BY + CZ + \ldots \]  

(Eq. S3)

LDA will not provide a unique boundary for data sets that are over-determined, meaning that the data set could not contain more variables than data elements\textsuperscript{14, 15}. In terms of MALDI IMS data, LDA would normally require that the number of pixels exceed the number of m/z values scanned within individual mass spectra. Unfortunately, this condition is rarely satisfied with MALDI IMS data sets as images would either need to contain thousands of pixels or the resolution within the mass spectra would need to decrease below the point of utility (e.g. binning mass spectra in increments of 250 m/z).

To review, PCA is considered a data reduction technique. Instead of constructing a mass spectrum using the recorded intensities of thousands of m/z values, PCA can construct an entire mass spectrum using far fewer independent variables (principal components) without changing the underlying content of the data as shown in Supplemental Figure S1. Because PCA dramatically reduces the number of independent variables, it can be used as a preliminary step prior to LDA, ensuring that data sets are not over-determined\textsuperscript{14, 15}.
When PCA is used prior to LDA, the mathematical equation of the linear boundary determined by LDA would have the form

\[(A \times PC_1) + (B \times PC_2) + (C \times PC_3) + (D \times PC_4) + \ldots\]  

(Eq. S4)

where A, B, C, and D are scalar coefficients and PC\(_1\), PC\(_2\), PC\(_3\), and PC\(_4\) are the relevant principal components (note that equation S4 would continue to include all relevant principal components of the data set). PCA offers another advantage prior to LDA in addition to the necessary data reduction step; noise within the mass spectra is discarded, improving the quality of the LDA boundary\(^{14}\).

As stated earlier, a single point within a PCA score plot represents an entire data spectrum. Mathematically, each data spectrum is represented as a linear combination of principal components:

\[\text{Spectrum} = (\text{Score}_{PC_1} \times PC_1) + (\text{Score}_{PC_2} \times PC_2) + (\text{Score}_{PC_3} \times PC_3) + (\text{Score}_{PC_4} \times PC_4) + \ldots\]  

(Eq. S5)

where specific score values generate the mass spectrum unique to a particular pixel within the MALDI IMS image. Much the same way that a recorded mass spectrum can be calculated using the linear combination of principal components and scores, a discriminant spectrum can be calculated using the linear combination of principal components and the scalar coefficients from the linear boundary (note the similarity in the mathematical formulas between equations S3, S4, and S5). Because the linear boundary drawn maximally separates two groups of data with different spectral features, the resulting discriminant spectrum will highlight spectral features that maximally differentiate the two groups of data\(^{14}\).
References