The Annals of Applied Statistics
2015, Vol. 9, No. 4, 1973–1996
DOI: 10.1214/15-AOAS870
© Institute of Mathematical Statistics, 2015

FEATURE EXTRACTION FOR PROTEOMICS IMAGING MASS SPECTROMETRY DATA

BY LYRON J. WINDERBAUM, INGE KOCH, OVE J. R. GUSTAFSSON, STEPHAN MEDING AND PETER HOFFMANN[1]

The University of Adelaide

Imaging mass spectrometry (IMS) has transformed proteomics by providing an avenue for collecting spatially distributed molecular data. Mass spectrometry data acquired with matrix assisted laser desorption ionization (MALDI) IMS consist of tens of thousands of spectra, measured at regular grid points across the surface of a tissue section. Unlike the more standard liquid chromatography mass spectrometry, MALDI-IMS preserves the spatial information inherent in the tissue.

Motivated by the need to differentiate cell populations and tissue types in MALDI-IMS data accurately and efficiently, we propose an integrated cluster and feature extraction approach for such data. We work with the derived binary data representing presence/absence of ions, as this is the essential information in the data. Our approach takes advantage of the spatial structure of the data in a noise removal and initial dimension reduction step and applies k-means clustering with the cosine distance to the high-dimensional binary data. The combined smoothing-clustering yields spatially localized clusters that clearly show the correspondence with cancer and various noncancerous tissue types.

Feature extraction of the high-dimensional binary data is accomplished with our difference in proportions of occurrence (DIPPS) approach which ranks the variables and selects a set of variables in a data-driven manner. We summarize the best variables in a single image that has a natural interpretation. Application of our method to data from patients with ovarian cancer shows good separation of tissue types and close agreement of our results with tissue types identified by pathologists.

1. Introduction. Mass spectrometry (MS) has become a versatile and powerful tool in proteomics for the analysis of complex biological systems, including the identification and quantification of proteins and peptides [Ong and Mann (2005)]. Many different technologies have been developed under the collective field of proteomics mass spectrometry [Aebersold and Mann (2003)]. The focus in this paper is the more recent development [see Aoki et al. (2007), Groseclose et al. (2006)] of matrix assisted laser desorption ionization (MALDI) imaging mass spectrometry

Received June 2014; revised August 2015.

[1]Supported in part by the Australian Research Council (ARC LP110100693), Bioplatforms Australia and the Government of South Australia.

Key words and phrases. Proteomics, mass spectrometry data, high-dimensional, binary data, MALDI-IMS, unsupervised feature extraction.

(IMS), also known as MALDI imaging, and, in particular, an analysis of MALDI-IMS data acquired from tissue samples of patients with ovarian cancer.

Unlike the more common 2D gel electrophoresis (2D-GE) and liquid chromatography (LC) based techniques in proteomics, MALDI-IMS preserves the spatial distribution inherent in the tissue; and the tens of thousands of spatially distributed spectra acquired from a single tissue sample in a MALDI-IMS experiment provide new challenges for statisticians and bioinformaticians as well as having the potential to lead to breakthroughs in biological research [see Casadonte and Caprioli (2011)]. We propose a combined cluster and feature extraction method for such data which exhibits cancer-specific variables whose protein associations can be inferred by parallel LC-MS experiments such as those of Meding et al. (2012).

Standard proteomics mass spectrometry methods such as 2D-GE and LC-MS have been described in the literature for some time; see Wasinger et al. (1995). We briefly explain LC-MS and important differences with MALDI-IMS in Section 2. For an overview and review of recent approaches in LC-MS, see America and Cordewener (2008). Statistical challenges of proteomics mass spectrometry data are outlined in Wu et al. (2003). The statistics and bioinformatics literature on the analysis of 2D-GE and LC-MS data is growing fast and covers a range of statistical methods. Testing and classification of such data are described in Morris (2012), Morris et al. (2005), Yu et al. (2006) and references therein. Other statistical approaches that have been proposed and applied to 2D-GE and LC-MS data include peak identification, alignment and feature selection [see Yu et al. (2006)], identification of proteins [see Yu et al. (2006) and Karpievitch et al. (2010)], wavelet-based methods [see Morris and Carroll (2006), America and Cordewener (2008), Du, Kibbe and Lin (2006) and references therein] and methods from survival analysis for the detection of differentially expressed proteins [see Tekwe, Carroll and Dabney (2012)].

Contrasting these developments in the analyses of 2D-GE and LC-MS data, the newer MALDI-IMS methods which have been introduced into routine research practice have not yet attracted as much attention in the statistics literature, although MALDI-IMS methods are covered in proteomics/mass spectrometry journals—see Alexandrov and Kobarg (2011), Alexandrov et al. (2010, 2013), Gessel, Norris and Caprioli (2014), Jones et al. (2012), Norris et al. (2007), Stone et al. (2012) and references therein. The potential of MALDI-IMS is described in Alexandrov and Kobarg (2011): "*IMS is one of the most promising innovative measurement techniques in biochemistry which has proven its potential in discovery of new drugs and cancer biomarkers*.... *IMS was used in numerous studies leading to understanding chemical composition and biological processes*.... *As for many modern biochemical techniques*, *in particular in proteomics*, *the development of computational methods for IMS is lagging behind the technological progress*." In addition to presenting our approach and analyses of MALDI-IMS data, we hope to motivate other statisticians and bioinformaticians to explore this exciting and promising new area and to develop novel statistical methods for the analysis of such data.

Related to our research are the papers by Alexandrov et al. (2010), Deininger et al. (2008) and Bonnel et al. (2011) who cluster their MALDI-IMS data using principal component analysis and hierarchical clustering, or Gaussian mixture models. Our proposal, outlined in Figure 1, differs from their research in a number of important aspects. Unlike these authors, we derive suitably binned binary data, which we describe in Section 3, instead of working with the raw or intensity data. Following [Koch (2013), Chapter 6], who demonstrates the success of using such binary data in finding biologically meaningful tissue clusters, we apply k-means

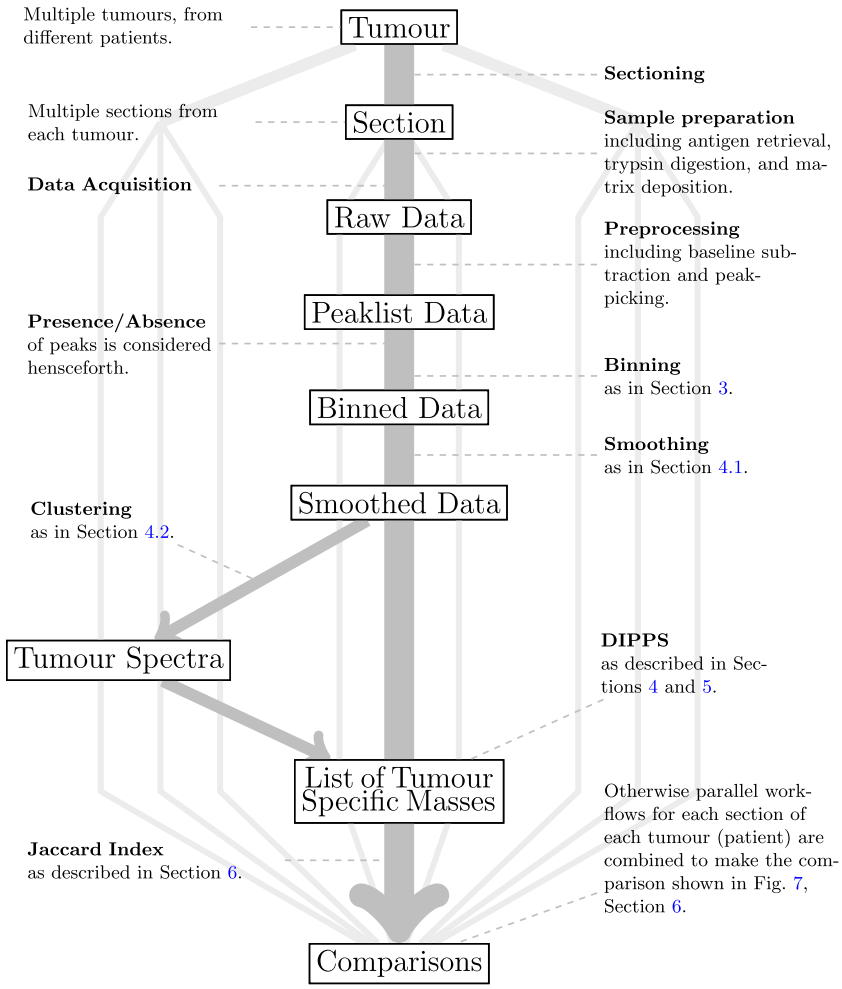

FIG. 1. *Data processing workflow for our proposed combined clustering-DIPPS approach. The background arrows (in light grey) represent parallel workflows for the three sections from each of the three patients, each leading to a "list of tumor specific masses." In the final step the lists are compared as described in Section 6.*

clustering to the binary data. Analysis of the binary data has the added advantage of being computationally more efficient. Furthermore, parallel computation is possible for all steps in the approach we propose. Our approach combines clustering with explicit feature extraction conceptually similar to the approach of Jones et al. (2011), although based on a different principle—our Difference in Proportions of Occurrence (DIPPS) statistic which ranks and selects the "best" variables in a data-driven way. Our use of binary data allows the feature extraction results to be visualized as a single heat map with a natural interpretation. The ability to visualize the selected features as a single easily interpretable image has not been a part of the above-mentioned papers and gives a significant advantage to our approach.

This paper is organized in the following way. We briefly describe relevant background on proteomics in Section 2, and discuss advantages of MALDI-IMS for the biological research fields. We describe MALDI-IMS data and how to derive the binned binary data in Section 3. Section 4 covers the spatial smoothing and clustering steps of our method. The DIPPS approach (including a feature extraction step) is described in Section 5. Finally, in Section 6 we apply our combined cluster and DIPPS approach in order to compare several patients by considering several datasets. The results discussed in Section 6 allow us to demonstrate how these data could be used to address biologically relevant questions such as the identification of potential tissue-specific protein markers and classification of patients based on their response to treatment.

2. Proteomics background. Ovarian cancers are virtually asymptomatic and, as a result, the vast majority of cases are detected when the disease has metastasized. For these patients, radical surgery and chemotherapy are often insufficient to address the disease adequately and many patients relapse. The combination of late-stage diagnosis and unsuccessful treatments makes ovarian cancer the most lethal gynecological cancer, with advanced stage patients exhibiting a five year survival rate of less than 30% [Jemal et al. (2011), Ricciardelli and Oehler (2009)]. The keys to addressing ovarian cancer will be as follows: increasing our understanding of the mechanisms driving cancer progression, identifying molecular markers which can predict treatment success and identifying new treatment targets. As proteins are key functional components of cells and tissues, determining protein distributions in cancer tissue represents a crucial step in addressing these key aims.

Proteins are synthesized within cells as linear amino acid sequences and folded into more complex 3D structures that determine function and intracellular location. The complete set of proteins which exist in a given cell, tissue or biological fluid, under defined conditions, is termed its proteome [Wilkins et al. (1996)]. Proteomes vary considerably between different cellular states and understanding these variations allows insight into the development and progression of cancer. Proteomics characterizes proteome changes using a combination of fractionation, identification and quantitation strategies. Proteomics will use either a top-down or bottom-up approach. Top-down approaches analyze intact proteins, whereas bottom-up

approaches use proteolytic enzymes (e.g., trypsin) to digest proteins into peptides prior to analysis. The data we present is on tryptic peptides, so our discussion is in the context of a bottom-up approach. Proteome fractionation can be achieved using gel electrophoresis [Gygi et al. (2000)] or liquid chromatography (LC), with LC being the predominant fractionation technique [Rogowska-Wrzesinska et al. (2013)]. LC makes use of columns to affinity-bind molecules to a stationary phase. The molecules are subsequently eluted over time with a changing gradient of mobile phase solvent. In a bottom-up LC experiment peptides in a hydrophilic mobile phase are bound to a hydrophobic stationary phase. The peptides are eluted using an increase in the percentage of hydrophobic solvent in the mobile phase. To characterize the fractioned peptides, the LC eluant is often directly coupled to an MS instrument (LC-MS).

MS instruments contain an ion source, mass analyzer and detector. So-called "soft" ionization sources such as electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) are favored in proteomics, as they prevent significant molecular fragmentation during ionization. LC-MS instruments usually employ ESI to produce gaseous ions for mass analysis. The mass-to-charge ratio (m/z) of these ions is measured by the mass analyzer and detector to produce a mass spectrum of measured intensity as a function of m/z. Peptides analyzed by LC-MS can be fragmented to produce spectra which correspond to amino acid composition. Identification of these spectra is attempted using search algorithms, such as MASCOT [see Koenig et al. (2008) and references therein], which match measured spectra to expected fragmentation spectra. The combination of LC-MS with bottom-up strategies is what allows proteomic studies to identify and quantify thousands of unique proteins in a given biological sample.

LC-MS suffers from two method-specific limitations:

1. Tissue samples are homogenized and solubilized, which removes all spatial information inherent in the tissue, and
2. An LC-MS run usually takes more than an hour, precluding the rapid ($\leq$1 day) analysis of large sample numbers ($\geq$20).

Given that tissues are a mix of different cell populations and their organization is directly related to their functions, point 1 above can be crucial. Typically, tissue structure is visualized by a pathologist using histological stains such as haematoxylin and eosin (H&E) followed by light microscopy, as shown in Figure 2(a). Such histological stains allow visualization of the spatial distribution of cellular morphology and can provide an understanding of the way in which the cellular morphology relates to cancer behavior and, ultimately, the cancer's effect on the patient. The spatial information is essential in these histological stains, and it is reasonable to assume it would be equally crucial in mass spectrometry. The loss of spatial information that occurs during sample preparation for LC-MS analysis therefore motivated the development of direct tissue analysis using MALDI-IMS

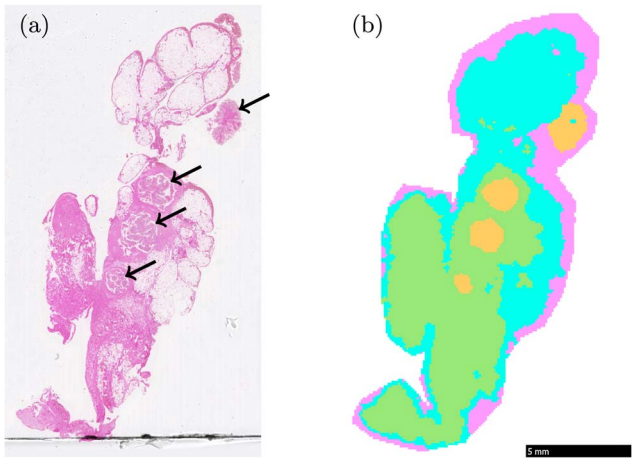

FIG. 2. *Left*, (a) *shows the H&E stained tissue section*, *arrows indicate the four cancer nodules*. *Right*, (b) *shows pixels plotted in their relative spatial locations*, *color identifies the cluster membership produced by* 4*-means clustering of the binary spatially smoothed data*.

[Cornett et al. (2007), Groseclose et al. (2008), Gustafsson et al. (2011)]. To prepare a sample for MALDI-IMS, a tissue block is thinly sectioned (2–10 μm thick slices) and mounted onto conductive microscopy slides. For analysis of peptides, the tissue section is coated with a homogeneous layer of proteolytic enzyme to digest endogenous proteins. The digest is then overlaid with a matrix compound which co-crystallizes with the tissue-derived peptides. A MALDI source is used to ionize these peptides directly from the tissue and facilitate the collection of a mass spectrum from each (x, y) position on a regular grid across the tissue section. An acquisition spacing of 20–250 μm is typical and balances the requirements for high quality MS data, practical data size and measurement time. In the datasets presented here, we used a spacing of 100 μm. The availability of spatial information in MALDI-IMS offers a unique perspective on tissue analysis, which is complementary to LC-MS.

In this context, the advantages of MALDI-IMS are as follows:

1. Hundreds of biological molecules can be measured in a single experiment in an untargeted manner, as compared to immunohistochemistry.
2. Tissue sections can be H&E stained post-acquisition of MALDI-IMS data [Deutskens, Yang and Caprioli (2011)] and spatial changes in protein abundance measured by MALDI-IMS can be compared to the histology visible in the H&E stain of the same tissue (see Figure 2). This makes the technique compatible with pathology.
3. MALDI-IMS provides a capacity for rapid interrogation of large sample numbers. This can be achieved using tissue microarrays (TMAs) [see Groseclose et al. (2008), Steurer et al. (2013)]; construction of TMAs involves the extraction

TABLE 1
Selection of peptide m/z values, the DIPPS score of the m/z bins containing them, and their inferred parent proteins. Peptide sequences and parent proteins were inferred by mass matching to concurrent LC-MS/MS analyses and validated by both in situ MALDI-MS/MS and immunohistochemistry as shown in Supplement A [*Winderbaum et al.* (2015a)]

LC-MS/MS mass $[M+H]^+$	**DIPPS statistic**	**UniProtKB/SwissProt entry name**	**Protein name**
1628.8015	0.662	ROA1_HUMAN	Heterogeneous nuclear ribonucleoprotein A1
2854.3884	0.910	K1C18_HUMAN	Keratin 18

of cylindrical cores from multiple tissue blocks and arrangement of these cores within a new block. A one-day sample preparation of sections from such a block allows an overnight MALDI-IMS experiment to collect data for $\geq$50 samples.

A caveat of MALDI-IMS analyses is that the proteome cannot be fractionated, and the masses measured using MALDI-IMS are not fragmented to provide sequence information. LC-MS based proteomics is required in order to infer parent proteins, as in Table 1. MALDI-IMS, however, places this information in a structural context and is therefore a crucial link between molecular composition and morphology.

MALDI-IMS therefore provides unprecedented capacities for both proteomics and pathology and, as a consequence, has important implications for research into human cancers [Gorzolka and Walch (2014)]. For example, ovarian cancers are known to be quite heterogeneous tissues [Deininger et al. (2008)]. Spatial analysis by MALDI-IMS allows this heterogeneity to be addressed explicitly [Gorzolka and Walch (2014)]. Furthermore, if TMAs are used, hundreds of patients can potentially be analyzed in a single day. MALDI-IMS provides the opportunity to (i) screen preprepared cancer samples rapidly, targeting regions of interest and, (ii) employ complementary LC-MS methods to characterize the tumor proteome of these tissues.

To exploit these capabilities in the future, it will be crucial to understand and distinguish differences between patients. As proof of the concept, in this paper we take data including multiple tissue types (e.g., tumor, adipose, connective) from full sections of tissue from each of three patients, and consider the problem of separating tumor-specific masses.

3. The data and binary binning. In Section 6 we consider a selection of datasets from three surgically excised ovarian cancers, each from a different patient; see Gustafsson (2012). An overview for our data analysis workflow is described in Figure 1. Thin cross-sections of tissue are obtained from each cancer,

and MALDI-IMS data are collected using a Bruker Ultraflex III and a m/z range of 1000–4500. Multiple sections are taken from each cancer, and we refer to the data collected from each section as a "dataset." Initially, we consider a single typical dataset (shown in Figure 2) and will refer to it as the "motivating dataset" throughout. A typical dataset might have 5000–100,000 spectra, the motivating dataset consists of 13,916. The motivating dataset will serve to illustrate our proposed method which we describe in Sections 4 and 5, completing the workflow illustrated in Figure 1. In Section 6 we consider the remaining datasets (represented by grey background arrows in Figure 1 and including the motivating dataset), three from each cancer, and compare datasets within and between patients. Although we have access to more datasets, we show only nine, as these suffice for comparisons both within and between patients while still illustrating our method concisely.

As discussed in Section 2, each dataset contains mass spectra collected at regularly spaced points on the surface of a thin tissue section. Each mass spectrum has annotation meta-data corresponding to the spatial (x, y) coordinates of its acquisition. The mass spectrum itself is a set of ion-intensity, m/z value pairs, and can be thought of as a discrete approximation of ion-intensity as a function of m/z. We will not explicitly address this functional aspect. Each mass spectrum typically contains between 5 and 200 intensity-peaks (local maxima), each corresponding to the m/z of a biological analyte such as a peptide or endogenous protein. A number of preprocessing steps are involved in extracting the peaks; we use methods available in proprietary software (flexAnalysis, Bruker Daltonik, http://www.bruker.com): smoothing (Gaussian kernels), baseline reduction (TopHat), and, finally, peak-picking (SNAP). The SNAP algorithm isolates monoisotopic peaks and defines significant peaks as those peaks with a signal-to-noise ratio of two or higher. Representing the data as peak lists significantly reduces the amount of data involved, often by as much as two orders of magnitude, and this can be very important due to the amount of data involved ($>$10 GB). We will be concerned only with these extracted peaks from here on.

Improved peak-picking is of interest for two reasons. First, peak-picking is currently the most computationally intensive step in our workflow, taking longer than the remainder of the data analysis combined. Conveniently, the Bruker peak-picking can be run simultaneously with the data acquisition, which typically takes 6–14 hours depending on the number of spectra acquired. Second, the choice of peak-picking algorithm will affect all subsequent data analysis and so improvements to the peak-picking algorithm would be expected to carry through and produce improved results downstream. Comparing the Bruker peak-picking that we have used with other existing methods and making improvements based on these comparisons is important, but is beyond the scope of this paper. All the code used to generate our results from the peaklist data is included in Supplement B [Winderbaum et al. (2015b)], and the peaklist data itself is in Supplement C [Winderbaum et al. (2015c)]. Software packages that implement existing peak-peaking algorithms and other relevant analysis tools include the R package

MALDIquant (http://strimmerlab.org/software/maldiquant/), the MATLAB bioinformatics toolbox (http://au.mathworks.com/help/bioinfo/index.html) and Bioconductor (http://www.bioconductor.org). For other mass-spectrometry based software resources, see http://www.ms-utils.org.

We bin the peaks by partitioning the m/z range into equal size intervals or bins, and identifying each peak with the bin it belongs too. Using a data-independent partitioning for the binning makes comparisons between datasets straightforward. The disadvantage is that peaks detected at very similar m/z values will be identified with different bins if a boundary happens to fall between them. In order to address this, we suggest running all analyses in tandem with shifted bin locations and combining the results in order to capture any analytes whose m/z is too close to a bin boundary. Using larger bin sizes also helps limit this effect. For the sake of brevity, we will omit the tandem analysis from the results, as it provides only a small improvement. We call the data after binning "binned intensity data," as in this form the variables correspond to the intensity (height) of peaks in particular m/z bins. We further reduce the "binned intensity data" to "binary binned data" where the variables (m/z bins) are binary (one/zero) valued—corresponding to presence/absence of peaks as described by Koch (2013), Example 6.12. The number of m/z bins in the binned data will vary depending on the choice of bin size. Our analyses are not sensitive to choice of bin size, provided the bin size chosen is within a reasonable range (0.05–2 m/z). We use an intermediate bin size of 0.25 m/z, which yields 5891 (nonempty) m/z bins in the motivating dataset. The principal effect that choice of bin size has is on the total number of m/z bins, and the number that are removed in the dimension reducing steps of our method—smaller bin sizes will result in more m/z bins initially, and more being removed, larger bin sizes result in fewer m/z bins initially, and fewer being removed.

The use of the binary data has a number of advantages, including separation of tissue types, computational efficiency, and allowing the use of the easily interpretable DIPPS approach we describe in Section 5. Extraneous variables such as matrix crystal morphology can have an adverse effect on the intensity measurements by adding noise [Garden and Sweedler (2000)]. Deininger et al. (2011) attempt to address these effects by normalization of their intensity values. One additional advantage of using the binary transformation suggested by [Koch (2013), Example 6.12] is that it circumvents the effects of such extraneous variables by avoiding direct use of the intensity values, and in this sense can be considered a data cleaning step. All subsequent analyses will concern only the binary data, as indicated in Figure 1, and we will refer to these simply as "the data" from here on. Similarly, we will refer to the variables in these data as "m/z bins," as we have done in this section.

4. Clustering spatially smoothed data. We propose a two-step method for separating spectra into groups corresponding to tissue types: first, a smoothing step which acts as a data cleaning and dimension reduction step, and incorporates the spatial information from MALDI-IMS into the data. Second, a clustering step.

4.1. *Smoothing step*. The spatial information available in MALDI-IMS data can be used to clean the data and remove m/z bins that are spatially dispersed. We incorporate this spatial information through a spatial smooth.

Let $\mathbb{X}$ be a $d \times n$ binary matrix; the rows of $\mathbb{X}$ correspond to m/z bins and are denoted $\mathbf{x}_{i\bullet}$, the columns of $\mathbb{X}$ correspond to spectra and are denoted $\mathbf{x}_{\bullet j}$, and the entries of $\mathbb{X}$ are denoted x_{ij}. These entries x_{ij} take the value one if peaks are present, and zero if no peaks are present in the m/z bin i in spectrum j. Let $0 \le \tau < \frac{1}{2}$ be a smoothing parameter and $\delta \ge 0$ a distance cutoff.

We iteratively update the values of $\mathbb{X}$. Let $\mathbb{X}^{(k)}$ denote the updated matrix at the kth iteration. Similarly, let $\mathbf{x}_{i\bullet}^{(k)}$, $\mathbf{x}_{\bullet j}^{(k)}$ and $x_{ij}^{(k)}$ denote the rows, columns and values of $\mathbb{X}^{(k)}$, respectively. At the kth iteration, the proportion of spectra, $T_{ij}^{(k)}$, in a spatial δ-neighborhood of the jth spectrum $\mathbf{x}_{\bullet j}^{(k)}$ whose values at the ith m/z bin $\mathbf{x}_{i\bullet}^{(k)}$ agree with $x_{ij}^{(k)}$ is

$$T_{ij}^{(k)} = \left\{ \left(1 - x_{ij}^{(k-1)}\right) + \left(2x_{ij}^{(k-1)} - 1\right)\left(\frac{\mathbf{x}_{i\bullet}^{(k-1)} \mathbf{d}_j^{\top} - x_{ij}^{(k-1)}}{\mathbf{1}_{1\times n}\mathbf{d}_j^{\top} - 1} \right) \right\}. \tag{1}$$

This proportion determines if the value $x_{ij}^{(k)}$ should be changed. In (1), $\mathbf{d}_j$ denotes a $1 \times n$ indicator vector with entry 1 if the corresponding indexed spectrum is in a δ-neighborhood of $\mathbf{x}_{\bullet j}$ and zero otherwise. If this proportion $T_{ij}^{(k)}$ is less than τ, we update the value $x_{ij}^{(k)}$ as in (2).

We generate the smoothed data by iteratively calculating the entries of $\mathbb{X}^{(k)}$:

$$x_{ij}^{(k)} = \begin{cases} x_{ij}^{(k-1)}, & \text{if } T_{ij}^{(k)} > \tau, \\ 1 - x_{ij}^{(k-1)}, & \text{if } T_{ij}^{(k)} \le \tau, \end{cases} \tag{2}$$

for $k = 1, 2, \ldots,$ starting with $\mathbb{X}^{(0)} = \mathbb{X}$. We stop when convergence is reached, that is, when $k = k^* = \min\{k : \mathbb{X}^{(k)} = \mathbb{X}^{(k-1)}\}$. The spatially smoothed data are $\mathbb{X}^{(k^*)}$.

Remarks on the smoothing process:

1. Without loss of generality, we let the distance between adjacent pixels be one. We choose $\delta = \sqrt{2}$ which results in a range 1 *Moore neighborhood*; see Gray (2003). This neighborhood is used in the cellular automata literature including Gardner (1970). It is worth noting that acquiring the range 1 Moore neighborhood by using the Euclidean distance and $\delta = \sqrt{2}$ on a regular grid is equivalent to using the Tchebychev distance, and $\delta = 1$.

2. The smoothing parameter τ defines the proportion of neighboring spectra needed to agree in order for a particular value to remain unchanged at any given step. Small values of τ smooth less ($\tau = 0$ leaves the data unmodified), while larger values smooth more. Results will not significantly change if τ is within the same $\frac{1}{8}$-wide interval, as changing τ within these intervals will affect spectra only on the

boundary of the acquisition region (spectra with less than 8 neighbors). The limit $\tau \to \frac{1}{2}$ results in maximum smoothing and is equivalent to the intuitive median smooth. In practice, the median smooth tends to yield over-smoothed data and often fails to converge. We chose an intermediate smoothing parameter, $\tau = \frac{1}{4}$, for these analyses. The values $\frac{1}{8}$ and $\frac{3}{8}$ could also be used, for less or more smoothing, respectively.

3. Alternative smoothing options include kernel methods [Wand and Jones (1995)] which apply to continuous data. These methods produce continuous values when applied to binary data, for which there is no clear interpretation. Our method produces binary smoothed data, maintaining the interpretability of the binary values.

4. At each smoothing iteration k, m/z bins are smoothed independently, and within each m/z bin all observations are smoothed simultaneously at each step. This means that it is possible to parallelize the smoothing algorithm, making efficient use of computational resources.

5. Our smoothing step plays a similar role in our approach to the combined two-step method of Alexandrov and Bartels (2013). Alexandrov and Bartels (2013) use first an edge-preserving smooth [Tomasi and Manduchi (1998)] and then a measure of spatial chaos to remove spatially chaotic images. Our method also removes spatially chaotic images by reducing them to empty. Improvements could potentially be achieved by combining the two approaches.

Bins that exhibit occurrence of peaks in a small number of spatially delocalized spectra or in almost all spectra constitute a large proportion of all m/z bins. These bins tend not to be relevant, as they are usually internal calibrants [Gustafsson et al. (2012)], errors, contaminants or tissue regions that are too small to be of interest due to the spatial (lateral) resolution used (100 μm). This last point could be improved by using a finer lateral resolution, as discussed by Schober et al. (2012). In these data, however, biological structures which are the same size as or smaller than the acquisition resolution, in this case 100 μm, will be removed by the smoothing. These m/z bins have zero variance after the smoothing step. Following the smoothing, these zero-variance m/z bins are removed, reducing the dimension of the data. The motivating dataset has 5891 m/z bins before the smoothing step. After the smoothing step 1022 of these m/z bins have nonzero variance and the remainder are discarded.

4.2. *Clustering step*. The second step in our approach concerns clustering of the spatially smoothed data. We use k-means clustering. Based on the information available from the histology, there are three broad tissue types present which could be labeled as fatty, connective and cancer tissue, respectively. Further, there are spectra that were acquired off-tissue. Thus, we perform k-means clustering with $k = 4$. We choose initial cluster centers at random from the sample, repeat this process 100 times and choose the clustering with minimum within-cluster sum of spectra-to-centroid distances.

k-means clustering of the binned intensity data with the default Euclidean distance does not lead to interpretable or spatially localized clusters [Koch (2013), Example 6.12]. In contrast, k-means clustering of the binned intensity data with the cosine distance, and of the binary data with the Euclidean or cosine distance, leads to clusters that correspond to the different tissue regions. Since there is no clear superiority of one distance over the other for the binary data, we continue with only the cosine distance, which has the added advantage of having become an established measure of closeness for high-dimensional data and associated consistency results [see Koch (2013), Sections 2.7, 13.3 and 13.4, and references therein].

For data with associated spatial information (such as MALDI-IMS data), it is natural to display the cluster membership in the form of cluster maps or cluster images: colored pixels at the (x, y) coordinates of the spectra which show the cluster membership of each spectrum using different colors to identify clusters. The H&E stained tissue cross-section and result of 4-means cluster analysis (by cosine distance) of the motivating dataset are shown in Figure 2(a) and (b), respectively. The cluster membership in Figure 2(b) corresponds well with tissue types as determined by the histology in Figure 2(a): cyan corresponds to fatty tissue, green to cancer-associated connective tissue and orange to four cancer nodules [indicated by arrows in Figure 2(a)]. The fourth cluster in pink corresponds well with off-tissue spectra, apart from a small amount of "bleed-out" from the cyan cluster possibly caused by nonspecific or "leakage" analytes. The correspondence between cluster results and histology demonstrates that the spatial smooth and cluster analysis isolate key molecular information which allows differentiation of tissue types by their mass spectra.

This correspondence between cluster results and histology, particularly for the cancer tissue-type, is important for the interpretation of results that follow, and so we took extra steps to validate its accuracy. A pathologist annotated the H&E stained tissue section for the motivating dataset [shown in Figure 2(a)], indicating regions of cancerous tissue. In order to avoid ambiguity in annotation, we then created an annotation subset identifying spectra whose origin is unambiguously cancerous tissue (omitting spectra from tissue regions of mixed tissue types, or tissue of ambiguous type). This allows us to be confident that spectra in this annotation subset are definitely from cancerous tissue, giving us a diagnostic measure of accuracy for our cancer cluster. The annotation subset we obtained for the motivating dataset contained 515 spectra, 499 (97%) of which where also contained in the 778 spectra of the orange cancer cluster of Figure 2(b).

5. The difference in proportions of occurrence (DIPPS) approach. In Section 4 we mention the good agreement between tissue types visible in the histology of Figure 2(a) and clusters of Figure 2(b). From the biological perspective it is of great interest to be able to quantify the differences between these tissue types in an easily interpretable way. At a mathematical level, a characterization of the differences between the tissue types translates into an identification of m/z bins that discriminate them. We propose the DIPPS approach for identifying discriminat-

ing m/z bins. Other methods for determining m/z values exist in the supervised learning literature, for example, support vector machines and PCA-based linear discriminant analysis. A comprehensive comparison of the DIPPS approach with these methods in determining the "best" and "correct number" of discriminating m/z values is needed, but such a comparison is beyond the scope of this paper. Instead we restrict attention to the new DIPPS approach which we first explain for an arbitrary subset of binary data, and then apply to the motivating dataset using spectra from the cancer cluster as the subset of interest, as distinguishing these spectra from the noncancer spectra is particularly important.

We define the DIPPS statistic for a fixed subset of data in (3), and show how this new statistic introduces a ranking of the m/z bins based on their ability to characterize the subset of interest. This DIPPS statistic leads to a natural heuristic, introduced in (4), for selecting a number of the ranked m/z bins, which we call *DIPPS features*, that "best" characterize the subset of interest. The extraction of these DIPPS features can be interpreted as a dimension reduction step. For data with spatial meta-data, such as MALDI-IMS data, we propose a way of displaying graphically the information obtained from the statistic in an easily interpretable summary image. These maps make this technique useful in exploratory analyses, as combining the DIPPS features into a single interpretable image allows for broad conclusions to be drawn quickly and easily. When the amount of data becomes large, the approach commonly used in proteomics, namely, of considering each m/z bin individually, is of limited use and the ability to produce a single image that summarizes many m/z bins becomes particularly useful.

Let $\mathcal{S}$ be a subset of observations (columns) of the data, $\mathbb{X}$. We let $\mathbf{p}(\mathcal{S})$ denote the mean of the observations in $\mathcal{S}$ and, similarly, let $\mathbf{p}(\mathcal{S}^c)$ denote the mean of the observations in its compliment $\mathcal{S}^c$. As the data are binary, the kth entry of the vector $\mathbf{p}(\mathcal{S})$ is the proportion of observations in $\mathcal{S}$ that take the value one for the kth m/z bin. The interpretation of binary data as the occurrence of an event—here existence of peaks—allows the mean $\mathbf{p}(\mathcal{S})$ to be interpreted as the proportions of occurrence (for each m/z bin) in $\mathcal{S}$. Considering occurrence (presence of peaks) in each m/z bin as a predictor of which spectra should be in $\mathcal{S}$ allows us to interpret the corresponding entry of $\mathbf{p}(\mathcal{S})$ as the "sensitivity" or true positive rate of this prediction. Similarly, each entry of the vector $\mathbf{1}_{d\times 1} - \mathbf{p}(\mathcal{S}^c)$ is the proportion of observations in $\mathcal{S}^c$ that take the value zero for the corresponding m/z bin, and, from the perspective of treating each m/z bin as a predictor for which spectra should be in $\mathcal{S}$, can be considered the "specificity" or true negative rate. In order to characterize $\mathcal{S}$, both sensitivity and specificity should be high. We sum these measures of sensitivity and specificity, and subtract one to give a range of $[-1, 1]$ and define the vector of DIPPS statistics $\mathfrak{d}$ for $\mathcal{S}$ as

$$\mathfrak{d}(\mathcal{S}) = \mathbf{p}(\mathcal{S}) - \mathbf{p}\big(\mathcal{S}^c\big). \tag{3}$$

For convenience of notation we omit the dependence on $\mathcal{S}$, and write $\mathfrak{d}(\mathcal{S}) = \mathfrak{d}$. We will similarly omit the $\mathcal{S}$ dependence for $\mathbf{t}_a$, n_a, $\mathbf{c}$ and a^* below, as we are treating

$\mathcal{S}$ as fixed. We use the entries of the vector $\mathfrak{d}$ to rank the m/z bins: the entry with the greatest value corresponds to the m/z bin that characterizes $\mathcal{S}$ best.

Next we determine the set of m/z bins that collectively best characterize $\mathcal{S}$, that is, the DIPPS features. We do this by finding a cutoff value in a data-driven way as follows. For $a > 0$ let $\mathbf{t}_a$ be the $d \times 1$ vector that takes the value one if the corresponding element of $\mathfrak{d}$ is $\geq a$, and takes the value zero otherwise. Let n_a be the number of entries in the vector $\mathbf{t}_a$ equal to one. Let $\mathbf{c}$ be the centroid of $\mathcal{S}$. For the cosine distance, D, this is the average of the normalized (to length one) vectors. We use the cutoff

$$a^* = \arg\min_a \{D(\mathbf{c}, \mathbf{t}_a)\}. \tag{4}$$

$\mathbf{t}_a$ is a binary "template" vector for representing observations in $\mathcal{S}$. The centroid $\mathbf{c}$ represents the "center" of observations in $\mathcal{S}$. In (4) we choose a^* such that $\mathbf{t}_{a^*}$ is as close to $\mathbf{c}$ as possible. The n_{a^*} DIPPS features are the m/z bins whose corresponding entry in $\mathbf{t}_{a^*}$ is one.

In the motivating dataset we are interested in characterizing the cancer spectra, and thus we choose $\mathcal{S}$ to be the set of spectra belonging to the orange cluster shown in Figure 2(b). The cutoff of (4) is $a^* = 0.126$ and results in $n_{a^*} = 70$ DIPPS features being selected from the 1022 m/z bins remaining after smoothing. For each spectrum $\mathbf{x}_{\bullet j}$, the sum of the DIPPS features $\mathbf{t}_{a^*}^{\mathsf{T}}\mathbf{x}_{\bullet j}$, can be visualized spatially in a "DIPPS map," as shown in Figure 3(b). We construct the DIPPS map pointwise at each (x, y) location. The value of the DIPPS map at each (x, y) location, or pixel, represents the number of DIPPS features exhibiting occurrence

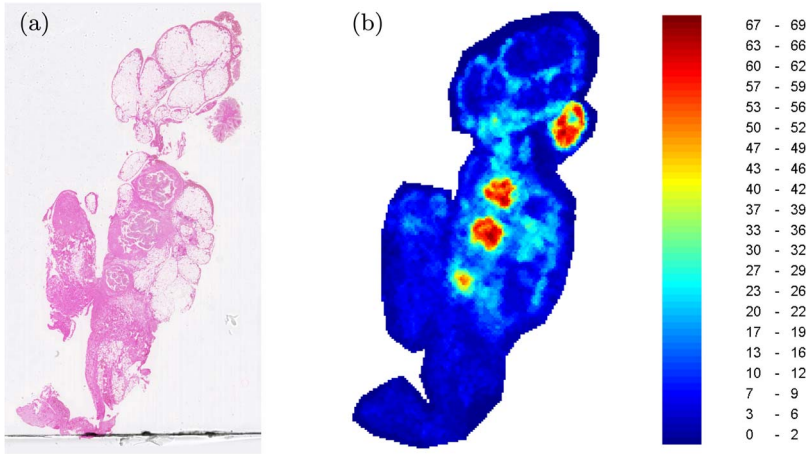

FIG. 3. *H&E stained tissue section* (a) *and DIPPS map* (b) *for the cancer cluster of the motivating dataset. The DIPPS map shows the sum of the* $n_{a^*} = 70$ m/z *bins with DIPPS score greater than* $a^* = 0.126$. *The legend relates the counts in heat colours*: *cold* (*blue*) *indicating spectra in which none of the selected 'DIPPS feature'* m/z *bins contain peaks*, *hot* (*red*) *indicating spectra in which all "DIPPS feature"* m/z *bins contain peaks*....

for the spectrum at (x, y). These counts are visualized in heat colours: cold (blue) indicating spectra in which none of the selected 'DIPPS feature' m/z bins contain peaks, hot (red) indicating spectra in which all 'DIPPS feature' m/z bins contain peaks. The ability to visualize results in a DIPPS map, which is easy to interpret, is attractive, as considering so many m/z bins individually can be time consuming and fail to provide a "big-picture" perspective.

The 70 DIPPS features characterize the cancer tissue, and thus deserve inspection. Follow-up analysis can be done to identify the peptides that these m/z bins correspond to, and to draw inference as to their parent proteins. To illustrate the biological significance of these m/z bins, we compared their m/z values to the previously published results of Gustafsson (2012). Two of the highly ranked m/z bins (listed in Table 1, along with inferred parent proteins) were previously identified as highly expressed in cancer by manual assessment of spatial distributions. The identifications were achieved through both mass matching to LC-MS/MS as well as in situ MS/MS [Gustafsson (2012)]. The identity and histological distribution of these analytes (corresponding to m/z bins in our data) were successfully validated using immunohistochemistry (see Supplement A [Winderbaum et al. (2015a)]). This confirms that the DIPPS approach can find features of known importance. Crucially, the DIPPS approach produces a list of characterizing m/z bins more rapidly and comprehensively than manual inspection of individual m/z bins.

A DIPPS map such as that shown in Figure 3(b) has an intuitive interpretation that the results of cluster analysis do not. The DIPPS map highlights gradations which reveal finer detail than is possible in cluster maps. We discuss this point further in Section 6. This visualization using DIPPS maps becomes increasingly important in exploratory analyses when the number of patients and datasets increases, as it quickly becomes infeasible to consider each of the selected m/z bins individually. The DIPPS approach also allows m/z bins crucial to cluster/tissue differentiation to be isolated and summarized. This selection of DIPPS features inherent in the DIPPS approach can also be thought of as a variable reduction step by reducing the data to 70 m/z bins. More importantly, it successfully separates tissue type-specific m/z bins, addressing the heterogeneity of the tissue. This facilitates the comparison of datasets, which is the focus of Section 6.

6. Application to multiple datasets. In this section we consider nine datasets, including the motivating dataset. Of these datasets, three are from each of three different patients which we will refer to as patients A, B and C, respectively. We will refer to the three datasets from patient A as A1, A2 and A3, and similarly for the datasets from patients B and C. Dataset A1 is the motivating dataset. Each of the three datasets arising from the same patient is acquired from thin (6 μm) tissue sections of a single surgically excised tissue. Because of this experimental setup, we expect the cluster and DIPPS maps of datasets from the same cancer to be similar in terms of the location of the cancer clusters and the selected m/z bins. We aim to separate within-patient from between-patient variability among the datasets.

The results of our analyses of the nine datasets are displayed in Figures 4, 5 and 6 for patients A, B and C, respectively. Each figure corresponds to one patient and shows the three datasets in rows. Each row consists of an H&E stain on the

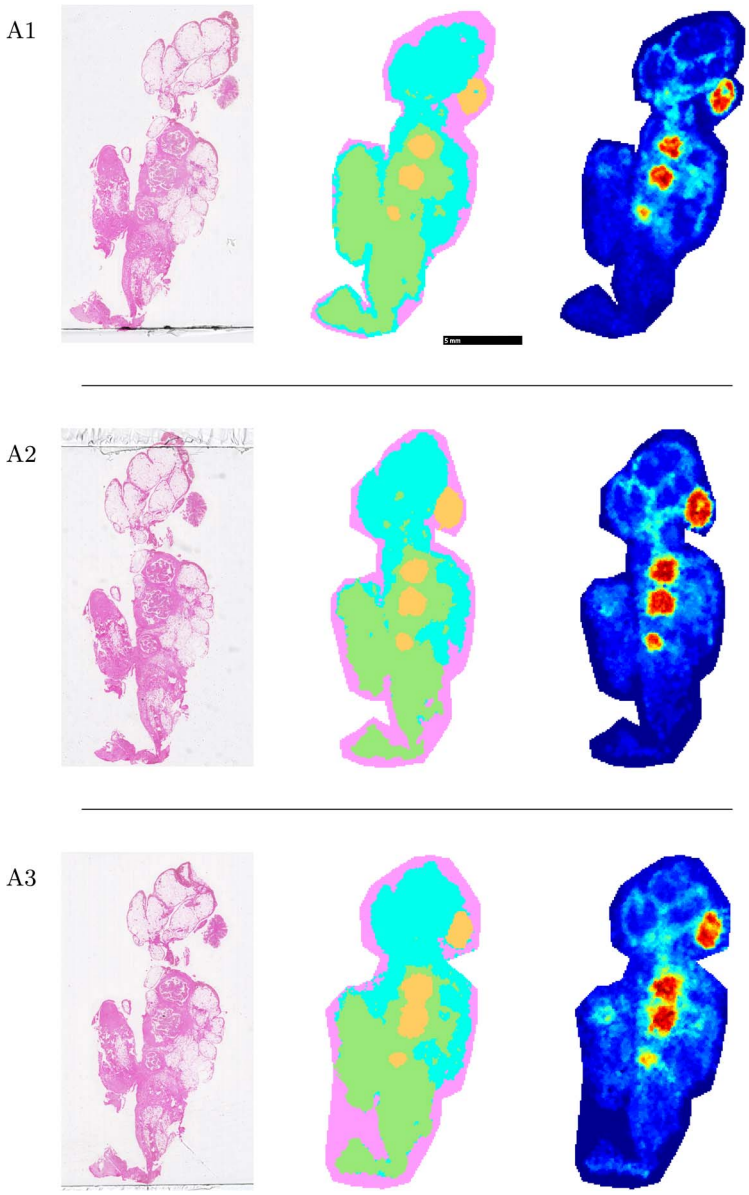

FIG. 4. *H&E stains* (*left column*), *cluster maps* (*center column*), *and DIPPS maps for the cancer cluster* (*right column*) *for the three datasets* A1, A2, *and* A3 (*each represented in a row*). *Note that* $n_{a^*} = 70, 45, 61$ m/z *bins are visualized in the DIPPS maps for the datasets* A1, A2, *and* A3, *respectively.*

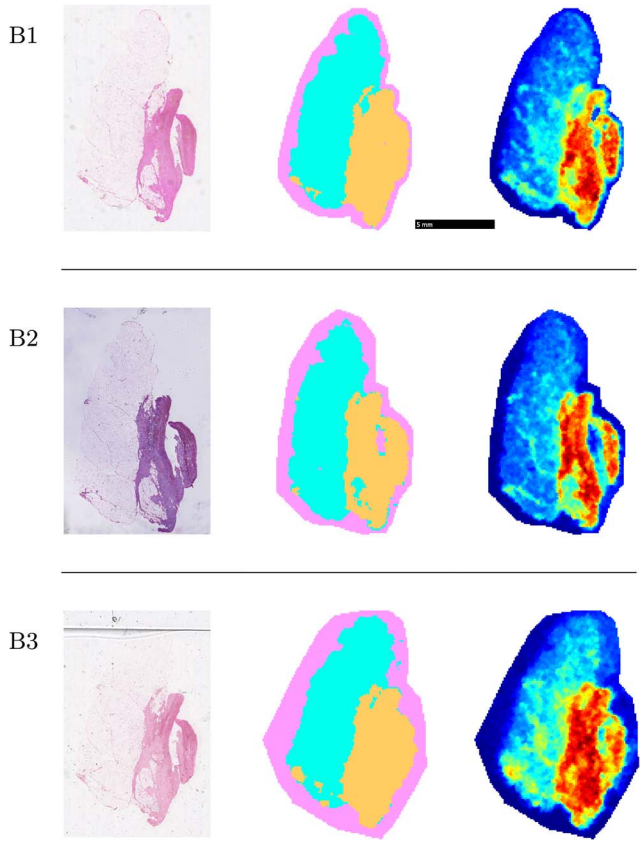

FIG. 5. *H&E stains* (*left column*), *cluster maps* (*center column*), *and DIPPS maps for the cancer cluster* (*right column*) *for the three datasets* B1, B2, *and* B3 (*each represented in a row*). *Note that* $n_{a^*} = 173, 117, 111$ m/z *bins are visualized in the DIPPS maps for the datasets* B1, B2, *and* B3, *respectively.*

left, a cluster map in the center and a DIPPS map for the cancer cluster on the right. The first row of Figure 4 repeats the results for the motivating dataset shown in Figures 2 and 3.

In all nine datasets, as visually judged by comparison with the H&E stained images, the clustering results correspond well with the tissue morphology. In patients B and C the connective tissue was more difficult to separate from the fatty tissue than in patient A, and so 3-means clustering was used instead of 4-means clustering. Disagreements between clustering results of datasets from the same patient serve to highlight the ability of the DIPPS maps to find and extract information in the data that is not available in the cluster maps, in a way that is remarkably robust to the clustering. For an example of this robustness property of the DIPPS maps, consider Figure 6—although the cluster map for dataset C2 shows a noticeable difference in the shape of its orange cancer cluster, the DIPPS maps show compar-

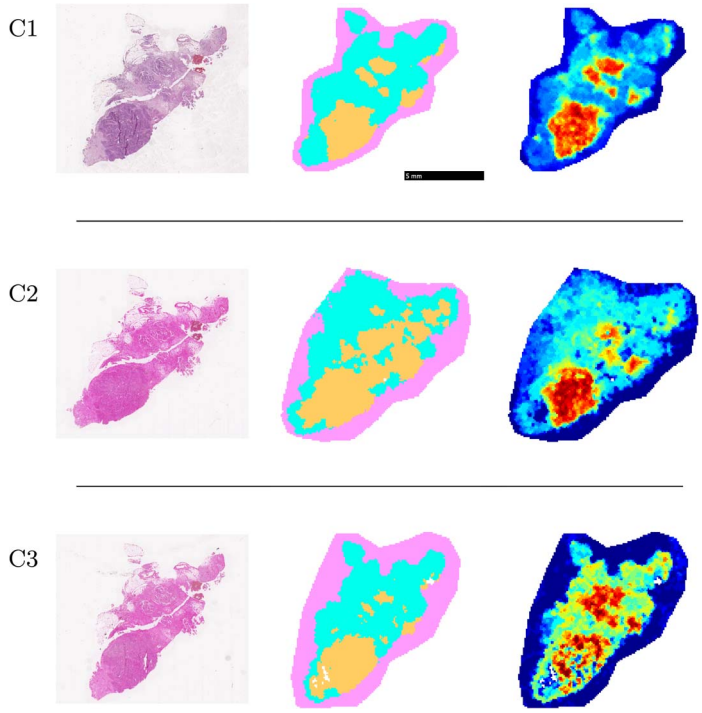

FIG. 6. *H&E stains (left column), cluster maps (center column), and DIPPS maps for the cancer cluster (right column) for the three datasets* C1, C2, *and* C3 *(each represented in a row). Note that* $n_{a^*} = 74, 38, 17$ m/z *bins are visualized in the DIPPS maps for the datasets* C1, C2, *and* C3, *respectively.*

atively consistent spatial distributions. Similarly, in Figure 5 datasets B1 and B3 show solid orange cancer clusters, failing to detect the clear vertical divide shown in the H&E stains, yet this divide is still apparent in the DIPPS maps.

The DIPPS maps are also more representative of the data, as they reveal gradual changes and fine detail that cannot be represented by the "hard" boundaries in a cluster map. For example, consider the bottom left region of tissue in Figure 5. The DIPPS maps highlight (subtly) a line following the bottom of the tissue corresponding to heterogeneous cancer-associated connective tissue. This is particularly interesting, as this "partial" highlighting in the DIPPS map indicates that there exists a subset of the selected m/z bins that exhibit presence in this region.

In addition to producing the DIPPS maps, the DIPPS approach yields a set of m/z bins that characterize the subset of the data in question. This set of DIPPS features can be used to compare tissue types across several datasets. We implement the DIPPS approach as described in Section 5 to identify DIPPS features for the cancer cluster in each of the nine datasets. To determine how similar these sets of characterizing m/z bins (DIPPS features) are, we use the Jaccard distance [Jaccard

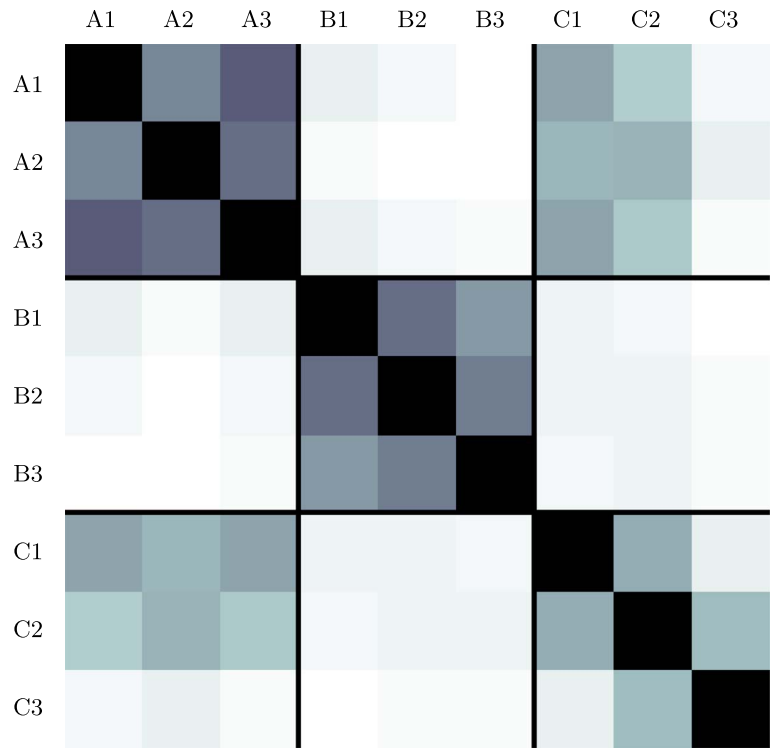

FIG. 7. *The* 9×9 *grid shows pairwise Jaccard distances between the sets of* m/z *bins characterizing the cancer clusters. Rows* (*and columns*) *correspond to the* 9 *datasets. The color of each pixel indicates the value of the Jaccard distance—white corresponding to a value of one, black to zero.*

(1901)]. For a pair of sets S_i, S_j, the Jaccard distance is

$$J(S_i, S_j) = 1 - \frac{|S_i \cap S_j|}{|S_i \cup S_j|}. \tag{5}$$

Figure 7 shows the pairwise Jaccard distances between the DIPPS features characterizing cancer. The "colors" range from darkest for a Jaccard distance of zero (indicating equality of sets) to white for a Jaccard distance of one, which indicates disjoint sets.

The block diagonal of Figure 7 is notably dark—this illustrates that differences within patients are smaller than differences between patients. This difference allows DIPPS features common to a particular patient to be separated from features that characterize cancer across multiple patients. Dataset C3 is an exception to this trend, as it shows notably less similarity to other datasets across the board (both within patient C and to datasets from other patients). Further inquiry into the raw data acquisition reveals that there was likely some problem at the sample preparation step for this slide, possibly in either the digestion or matrix deposition, as

fewer peaks are observed from spectra in dataset C3 in general when compared to typical spectra from any of the other datasets. In this way, our proposed exploratory analyses are also capable of alerting quality control issues, which would not have been obvious from the clustering results alone.

There are a number of DIPPS features common to all 3 patients, including the m/z bin centered at 1628.75, mentioned in Section 5. The peptide sequence and inferred parent protein of this feature (listed in Table 1) have been validated by in situ MS/MS and immunohistochemistry (shown in Supplement A [Winderbaum et al. (2015a)]), respectively, indicating that the protein from which this peptide derives is highly expressed in the cancer tissue of these patients. This protein could therefore be investigated further as a marker for ovarian cancer in a larger patient cohort.

Patient specific DIPPS features could be further investigated for their ability to classify cancers according to clinical or diagnostic criteria such as response to treatment. There appears to be a difference between patients A/C and B, as can be seen in Figure 7. This difference could be a consequence of the relatively larger number of DIPPS features identified in patient B datasets. The discrepancy in the number of selected m/z bins between patients A/C and B could be explained by the large amount of necrotic tissue in the patient B sample. Mass spectra from necrotic tissue are expected to be vastly different from those of other cancer tissue, which would indicate that the patient-specific m/z bins in patient B are likely to be markers for necrotic tissue.

Other tissue types could be considered (fatty and connective tissues), and a figure similar to Figure 7 could be produced for each of them. Such figures do not show a notably darker block-diagonal in the way that Figure 7 does. This tells us that the main differences between patients are in the patients' cancer, rather than in their other tissues, and reinforces how crucial it is to address the heterogeneity of these data by separating tissue types before conducting comparisons between patients.

7. Conclusion. This paper proposes an integrated approach to clustering and feature extraction for spatially distributed high-dimensional data. This approach is based on our difference in proportions (DIPPS) statistic and includes novel visualizations which enhance the cluster maps. For the MALDI-IMS cancer data, these maps have a natural interpretation in terms of the features that characterize cancer tissue. Application of our approach to different datasets from a number of patients allowed us to differentiate within-patient variability from between-patient variability.

In proteomics, the ability to automate feature extraction and to present these features as DIPPS maps provides an opportunity for holistic appraisal of MALDI-IMS data. This is crucial due to the size and number of such datasets and their high-dimensional nature. By isolating features important to specific tissue types and reporting similarities across patients, it will be easier to identify m/z bins for

further validation as tissue markers and to build models for addressing clinical questions such as predicting chemotherapy response and patient survival.

Acknowledgments. The authors gratefully acknowledge the histology annotation assistance provided by Andrew Ruszkiewicz (SA Pathology, Adelaide, South Australia). The authors thank the reviewers and the Editor for helpful comments which improved the paper.

SUPPLEMENTARY MATERIAL

Supplement A: Immunihistochemical Validation (DOI: 10.1214/15-AOAS870SUPPA; .zip). Optical images of immunohistochemical (IHC) tissue stains, validating three proteins as cancer-specific, including the two inferred parent proteins of Table 1. Top row are patient A replicates, bottom row patient C replicates.

Supplement B: Source Code (DOI: 10.1214/15-AOAS870SUPPB; .zip). Source code including cache and intermediate data files capable of reproducing all analyses up to and including compiling this document. Computations where done in `MATLAB`, and results compiled in LaTeX using the `R` package `knitr`.

Supplement C: Peaklist Data (DOI: 10.1214/15-AOAS870SUPPC; .zip). Raw peaklist data, used to generate the intermediate data files in Supplement B [Winderbaum et al. (2015b)].

REFERENCES

AEBERSOLD, R. and MANN, M. (2003). Mass spectrometry-based proteomics. *Nature* **422** 198–207.

ALEXANDROV, T. and BARTELS, A. (2013). Testing for presence of known and unknown molecules in imaging mass spectrometry. *Bioinformatics* **29** 2335–2342.

ALEXANDROV, T. and KOBARG, J. H. (2011). Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering. *Bioinformatics* **13** i230–i238.

ALEXANDROV, T., BECKER, M., DEININGER, S.-O., ERNST, G., WEHDER, L., GRASMAIR, M., VON EGGELING, F., THIELE, H. and MAASS, P. (2010). Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering. *J. Proteome Res.* **9** 6535–6546.

ALEXANDROV, T., CHERNYAVSKY, I., BECKER, M., VON EGGELING, F. and NIKOLENKO, S. (2013). Analysis and interpretation of imaging mass spectrometry data by clustering mass-to-charge images according to their spatial similarity. *Analytical Chemistry* **85** 11189–11195.

AMERICA, A. H. and CORDEWENER, J. H. (2008). Comparative LC-MS: A landscape of peaks and valleys. *Proteomics* **8** 731–749.

AOKI, Y., TOYAMA, A., SHIMADA, T., SUGITA, T., AOKI, C., UMINO, Y., SUZUKI, A., AOKI, D., DAIGO, Y., NAKAMURA, Y. et al. (2007). A novel method for analyzing formalin-fixed paraffin embedded (FFPE) tissue sections by mass spectrometry imaging. *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences* **83** 205–214.

BONNEL, D., LONGUESPEE, R., FRANCK, J., ROUDBARAKI, M., GOSSET, P., DAY, R., SALZET, M. and FOURNIER, I. (2011). Multivariate analyses for biomarkers hunting and validation through on-tissue bottom-up or in-source decay in MALDI-MSI: Application to prostate cancer. *Anal. Bioanal. Chem.* **401** 149–165.

CASADONTE, R. and CAPRIOLI, R. M. (2011). Proteomic analysis of formalin-fixed paraffin-embedded tissue by MALDI imaging mass spectrometry. *Nat. Protoc.* **6** 1695–1709.

CORNETT, D. S., REYZER, M. L., CHAURAND, P. and CAPRIOLI, R. M. (2007). MALDI imaging mass spectrometry: Molecular snapshots of biochemical systems. *Nat. Methods* **4** 828–833.

DEININGER, S.-O., EBERT, M. P., FÜTTERER, A., GERHARD, M. and RÖCKEN, C. (2008). MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. *J. Proteome Res.* **7** 5230–5236. PMID: 19367705.

DEININGER, S.-O., CORNETT, D. S., PAAPE, R., BECKER, M., PINEAU, C., RAUSER, S., WALCH, A. and WOLSKI, E. (2011). Normalization in MALDI-TOF imaging datasets of proteins: Practical considerations. *Anal. Bioanal. Chem.* **401** 167–181.

DEUTSKENS, F., YANG, J. and CAPRIOLI, R. M. (2011). High spatial resolution imaging mass spectrometry and classical histology on a single tissue section. *J. Mass Spectrom.* **46** 568–571.

DU, P., KIBBE, W. A. and LIN, S. M. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* **22** 2059–2065.

GARDEN, R. W. and SWEEDLER, J. V. (2000). Heterogeneity within MALDI samples as revealed by mass spectrometric imaging. *Analytical Chemistry* **72** 30–36.

GARDNER, M. (1970). Mathematical games: The fantastic combinations of John Conway's new solitaire game "life". *Scientific American* **223** 120–123.

GESSEL, M. M., NORRIS, J. L. and CAPRIOLI, R. M. (2014). MALDI imaging mass spectrometry: Spatial molecular analysis to enable a new age of discovery. *Journal of Proteomics* **107** 71–82. Special Issue: 20 in memory of Vitaliano Pallini.

GORZOLKA, K. and WALCH, A. (2014). November. MALDI mass spectrometry imaging of formalin-fixed paraffin-embedded tissues in clinical research. *Histology and Histopathology* **29** 1365–1376.

GRAY, L. (2003). A mathematician looks at S. Wolfram's new kind of science. *Notices Amer. Math. Soc.* **50** 200–211. MR1951106

GROSECLOSE, M. R., ANDERSSON, M., HARDESTY, W. M. and CAPRIOLI, R. M. (2006). Identification of proteins directly from tissue: In situ tryptic digestions coupled with imaging mass spectrometry. *J. Mass Spectrom.* **42** 254–262.

GROSECLOSE, M. R., MASSION, P. P., CHAURAND, P. and CAPRIOLI, R. M. (2008). High-throughput proteomic analysis of formalin-fixed paraffin-embedded tissue microarrays using maldi imaging mass spectrometry. *Proteomics* **8** 3715–3724.

GUSTAFSSON, O. J. R. (2012). Molecular characterization of metastatic ovarian cancer by MALDI imaging mass spectrometry. Ph.D. thesis, School of Molecular and Biomedical Science, Univ. Adelaide.

GUSTAFSSON, J. O. R., OEHLER, M. K., RUSZKIEWICZ, A., MCCOLL, S. R. and HOFFMANN, P. (2011). MALDI imaging mass spectrometry (MALDI-IMS)—application of spatial proteomics for ovarian cancer classification and diagnosis. *Int. J. Mol. Sci.* **12** 773–794.

GUSTAFSSON, J. O., EDDES, J. S., MEDING, S., KOUDELKA, T., OEHLER, M. K., MCCOLL, S. R. and HOFFMANN, P. (2012). Internal calibrants allow high accuracy peptide matching between MALDI imaging MS and LC-MS/MS. *Journal of Proteomics* **75** 5093–5105. Special Issue: Imaging Mass Spectrometry: A Users Guide to a New Technique for Biological and Biomedical Research.

GYGI, S. P., CORTHALS, G. L., ZHANG, Y., ROCHON, Y. and AEBERSOLD, R. (2000). Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci. USA* **97** 9390–9395.

JACCARD, P. (1901). Distribution de la Flore Alpine: Dans le Bassin des dranses et dans quelques régions voisines. Rouge.

JEMAL, A., BRAY, F., CENTER, M. M., FERLAY, J., WARD, E. and FORMAN, D. (2011). Global cancer statistics. *CA*: *A Cancer Journal for Clinicians* **61** 69–90.

JONES, E. A., VAN REMOORTERE, A., VAN ZEIJL, R. J., HOGENDOORN, P. C., BOVÉE, J. V., DEELDER, A. M. and MCDONNELL, L. A. (2011). Multiple statistical analysis techniques corroborate intratumor heterogeneity in imaging mass spectrometry datasets of myxofibrosarcoma. *PloS One* **6** e24913.

JONES, E. A., DEININGER, S.-O., HOGENDOORN, P. C., DEELDER, A. M. and MCDONNELL, L. A. (2012). Imaging mass spectrometry statistical analysis. *Journal of Proteomics* **75** 4962–4989. Special Issue: Imaging Mass Spectrometry: A User's Guide to a New Technique for Biological and Biomedical Research.

KARPIEVITCH, Y. V., POLPITIYA, A. D., ANDERSON, G. A., SMITH, R. D. and DABNEY, A. R. (2010). Liquid chromatography mass spectrometry-based proteomics: Biological and technological aspects. *Ann. Appl. Stat.* **4** 1797–1823.

KOCH, I. (2013). *Analysis of Multivariate and High-Dimensional Data*. Cambridge Univ. Press, New York. MR3154467

KOENIG, T., MENZE, B. H., KIRCHNER, M., MONIGATTI, F., PARKER, K. C., PATTERSON, T., STEEN, J. J., HAMPRECHT, F. A. and STEEN, H. (2008). Robust prediction of the mascot score for an improved quality assessment in mass spectrometric proteomics. *J. Proteome Res.* **7** 3708–3717.

MEDING, S., MARTIN, K., GUSTAFSSON, O. J., EDDES, J. S., HACK, S., OEHLER, M. K. and HOFFMANN, P. (2012). Tryptic peptide reference data sets for MALDI imaging mass spectrometry on formalin-fixed ovarian cancer tissues. *J. Proteome Res.* **12** 308–315.

MORRIS, J. S. (2012). Statistical methods for proteomic biomarker discovery based on feature extraction or functional modeling approaches. *Stat. Interface* **5** 117–135. MR2896986

MORRIS, J. S. and CARROLL, R. J. (2006). Wavelet-based functional mixed models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 179–199. MR2188981

MORRIS, J. S., COOMBES, K. R., KOOMEN, J., BAGGERLY, K. A. and KOBAYASHI, R. (2005). Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics* **21** 1764–1775.

NORRIS, J. L., CORNETT, D. S., MOBLEY, J. A., ANDERSSON, M., SEELEY, E. H., CHAURAND, P. and CAPRIOLI, R. M. (2007). Processing MALDI mass spectra to improve mass spectral direct tissue analysis. *Int. J. Mass Spectrom. Ion Phys.* **260** 212–221.

ONG, S.-E. and MANN, M. (2005). Mass spectrometry-based proteomics turns quantitative. *Nature Chemical Biology* **1** 252–262.

RICCIARDELLI, C. and OEHLER, M. K. (2009). Diverse molecular pathways in ovarian cancer and their clinical significance. *Maturitas* **62** 270–275.

ROGOWSKA-WRZESINSKA, A., LE BIHAN, M.-C., THAYSEN-ANDERSEN, M. and ROEPSTORFF, P. (2013). 2D gels still have a niche in proteomics. *Journal of Proteomics* **88** 4–13.

SCHOBER, Y., GUENTHER, S., SPENGLER, B. and RÖMPP, A. (2012). Single cell matrix-assisted laser desorption/ionization mass spectrometry imaging. *Analytical Chemistry* **84** 6293–6297.

STEURER, S., BORKOWSKI, C., ODINGA, S., BUCHHOLZ, M., KOOP, C., HULAND, H., BECKER, M., WITT, M., TREDE, D., OMIDI, M. et al. (2013). MALDI mass spectrometric imaging based identification of clinically relevant signals in prostate cancer using large-scale tissue microarrays. *Int. J. Cancer* **133** 920–928.

STONE, G., CLIFFORD, D., GUSTAFSSON, J. O., MCCOLL, S. R. and HOFFMANN, P. (2012). Visualisation in imaging mass spectrometry using the minimum noise fraction transform. *BMC Research Notes* **5** 419.

TEKWE, C. D., CARROLL, R. J. and DABNEY, A. R. (2012). Application of survival analysis methodology to the quantitative analysis of LC-MS proteomics data. *Bioinformatics* **28** 1998–2003.

TOMASI, C. and MANDUCHI, R. (1998). Bilateral filtering for gray and color images. 839–846, cited by 2167.

WAND, M. P. and JONES, M. C. (1995). *Kernel Smoothing*. Chapman & Hall, London. MR1319818

WASINGER, V. C., CORDWELL, S. J., CERPA-POLJAK, A., YAN, J. X., GOOLEY, A. A., WILKINS, M. R., DUNCAN, M. W., HARRIS, R., WILLIAMS, K. L. and HUMPHERY-SMITH, I. (1995). Progress with gene-product mapping of the mollicutes: Mycoplasma genitalium. *Electrophoresis* **16** 1090–1094.

WILKINS, M. R., PASQUALI, C., APPEL, R. D., OU, K., GOLAZ, O., SANCHEZ, J.-C., YAN, J. X., GOOLEY, A. A., HUGHES, G., HUMPHERY-SMITH, I. et al. (1996). From proteins to proteomes: Large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Nature Biotechnology* **14** 61–65.

WINDERBAUM, L. J., KOCH, I., GUSTAFSSON, O., MEDING, S. and HOFFMANN, P. (2015a). Supplement to "Feature extraction for proteomics imaging mass spectrometry data." DOI:10.1214/15-AOAS870SUPPA.

WINDERBAUM, L. J., KOCH, I., GUSTAFSSON, O., MEDING, S. and HOFFMANN, P. (2015b). Supplement to "Feature extraction for proteomics imaging mass spectrometry data." DOI:10.1214/15-AOAS870SUPPB.

WINDERBAUM, L. J., KOCH, I., GUSTAFSSON, O., MEDING, S. and HOFFMANN, P. (2015c). Supplement to "Feature extraction for proteomics imaging mass spectrometry data." DOI:10.1214/15-AOAS870SUPPC.

WU, B., ABBOTT, T., FISHMAN, D., MCMURRAY, W., MOR, G., STONE, K., WARD, D., WILLIAMS, K. and ZHAO, H. (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* **19** 1636–1643.

YU, W., WU, B., HUANG, T., LI, X., WILLIAMS, K. and ZHAO, H. (2006). Statistical methods in proteomics. In *Springer Handbook of Engineering Statistics* 623–638. Springer, Berlin.

SCHOOL OF MATHEMATICAL SCIENCES
THE UNIVERSITY OF ADELAIDE
ADELAIDE SA 5005
AUSTRALIA
E-MAIL: lyron.winderbaum@student.adelaide.edu.au
inge.koch@adelaide.edu.au