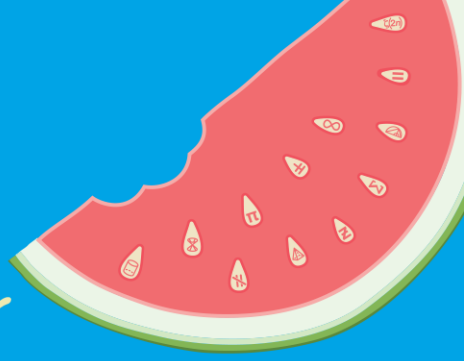
AMSI VACATION RESEARCH SCHOLARSHIPS 2021–22

Get a taste for Research this Summer



Statistical Analysis of Proteomic Mass Spectrometry Imaging Data

Sophie Giraudo

Supervised by Prof. Inge Koch

The University of Western Australia

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Abstract

Endometrial cancer is the most common gynaecological cancer and is both difficult to stage and has very low survival rates in later stages of the disease. Treatments currently include invasive and often unnecessary procedures that result in significant complications. It is therefore desirable to determine a method of classification using tissue markers to better stage the disease. This project uses k-means clustering and principal component analysis to determine differences in protein biomarkers between patients with early stage localised endometrial cancer, and those with late stage metastatic cancer. This analysis is an exploratory step in determining a classification method between patients with lymph node metastasis and those without.

1 Introduction

Endometrial cancer is the most common gynaecological cancer in Australia, and the most important factor for survival is the presence or absence of lymph node metastasis (LNM) (Winderbaum et al. 2016). Patients with a localised disease as seen in (a) and (b) in Figure 1 have a 96% 5-year survival rate, whereas those with a metastatic disease (Stages 3C, 4A, and 4B in Figure 1) have a 5-year survival rate of 17% (Rungruang and Olawaiye 2012).

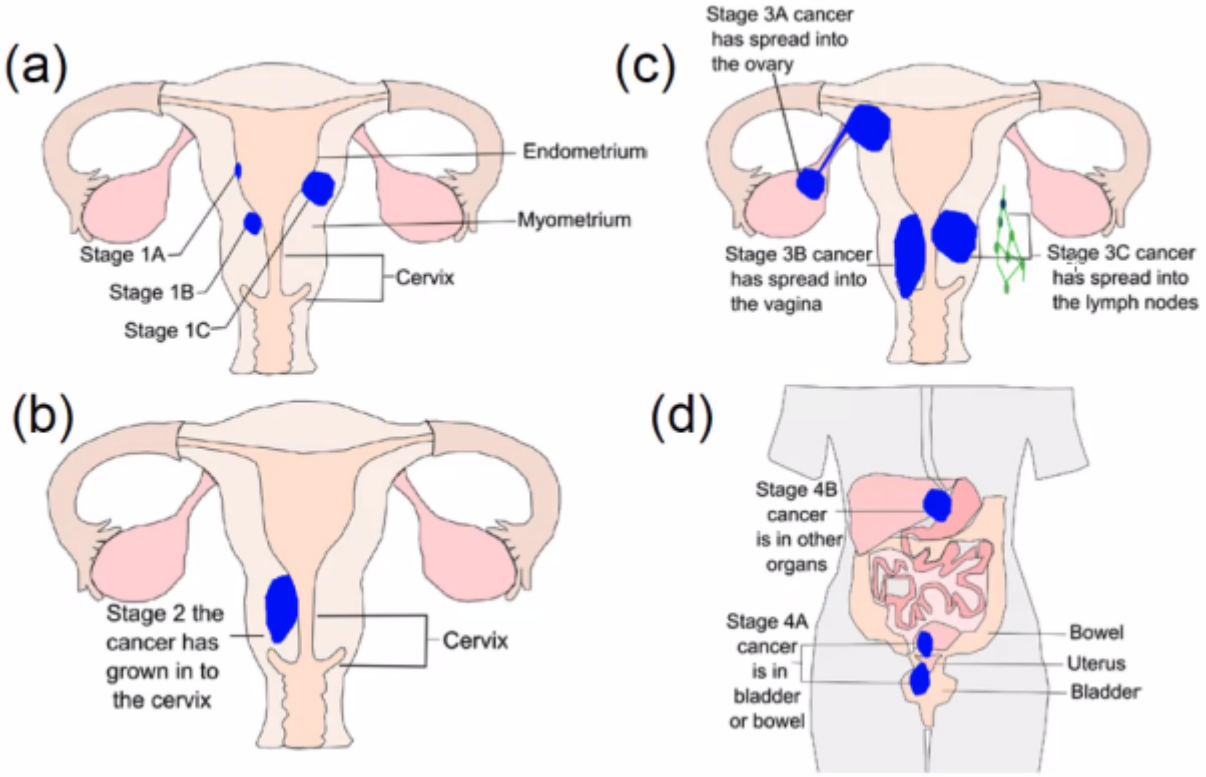


Figure 1: Stages of endometrial cancer.

Accurate staging of endometrial cancer is poor, and though LNM is only confirmed in approximately 15% of cases, most patients receive radical treatment including the removal of pelvic lymph nodes, resulting in

serious complications for nearly 38% of patients (Todo et al. 2010; Jacques et al. 1998). As such, a method of classification that uses predictive tissue markers of metastasis would benefit stage I and II patients by avoiding unnecessary and invasive procedures.

The first step of this classification is to determine if there are any differences in tissue markers of the cancers between LNM positive patients and LNM negative patients. Mittal et al. (2016) hypothesised and begun researching the potential to differentiate between metastatic and non-metastatic endometrial cancers at the proteome level through the use of matrix assisted laser desorption ionization mass spectrometry imaging (MALDI-MSI).

Previous work and this work considers data made available to us by Professor Peter Hoffman's lab at the University of South Australia. This data is collected from 185 patients, 34 of which are LNM positive, 140 are LNM negative, and 11 have an unknown diagnosis. For each patient, hundreds of mass spectra on more than 4500 mass bins are obtained using MALDI-MSI. Where previous work uses only the average spectrum for each patient, this work uses explicitly all mass spectra for each patient. The averaging for each patient may smooth out crucial information and differences between LNM positive and negative patients, and is therefore less likely to lead to accurate discrimination of the two groups.

Here, we cluster the spectra of each individual patient, the group of LNM positive patients, and the group of LNM negative patients using k-means clustering. We will then compare each patient's cluster pattern with its groups cluster pattern to potentially define a suitable measure of cluster similarity. We also aim to determine whether the data from patients in the same LNM group are more similar than the patterns of patients from different LNM groups. This clustering is an initial and exploratory study to gain valuable insight into differences between the groups. These differences may be used in future work to determine a method for classification. We use principal component analysis (PCA) to determine points of variance along the mass spectra.

1.1 Statement of Authorship

All of the statistical techniques used for this report are well established. PCA and K-means are university level topics of which I learnt in my undergraduate studies. The results and this report are all my own work.

I would like to acknowledge Inge Koch for her guidance on both appropriate statistical techniques, but also helping me develop a strong understanding of the background knowledge, as well as Peter Hoffman and his lab for providing the data and also assisting in my understanding of the problem.

2 Data

The data set used in this project contains protein biomarkers of endometrial cancers of 185 patients. 34 of these patients are LNM positive, 140 LNM negative, and 11 patients have an unknown diagnosis. For each patient, the data consists of 1 to 3 tissue samples that are then split into pixels in 2D space. At each of these pixels, MALDI-MSI captures the intensity spectrum as a function of mass-to-charge (m/z) ratios. These (m/z,

intensity) pairs are captured in the range 800-4000m/z, resulting in 171,157 pairs for each pixel. Figure 2 shows a plot of the first 100 of these pairs for a single pixel from a tissue sample.

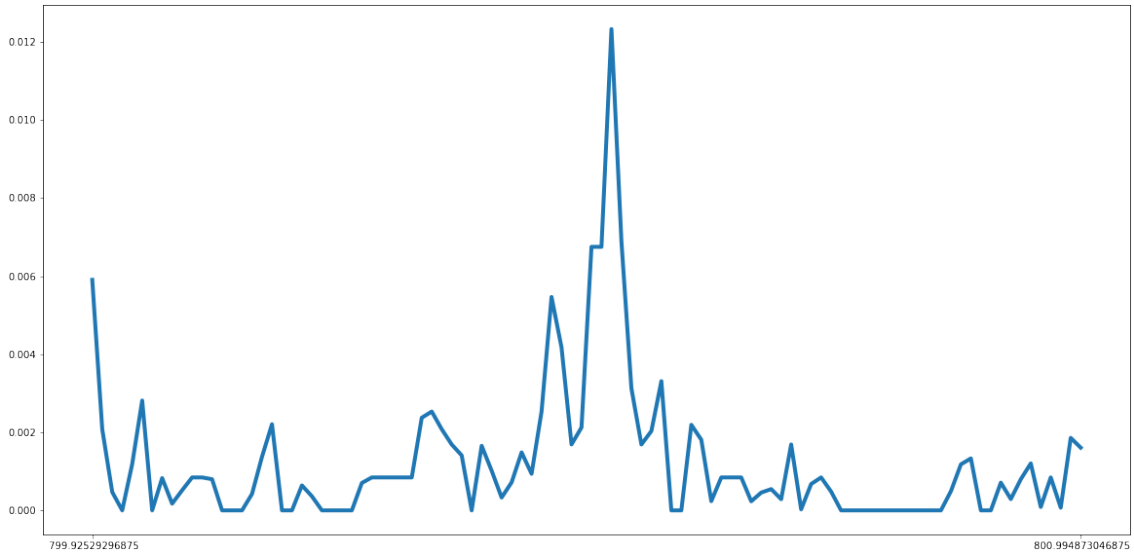


Figure 2: Beginning of the mass spectrum of a single pixel of a tissue sample from a patient that is LNM negative (patient 158).

Due to the size of the data, we consider only 8 patients who are LNM positive, and 8 who are LNM negative. This smaller subset of the data allows for patterns to still be identified, while having faster computation time.

3 Statistical techniques

3.1 *k*-Means clustering

To find a cluster pattern for each patient's spectra we use k-means clustering. k-Means clustering is an explorative technique that divides the data into k mathematically similar groups or clusters. We are required to specify the number of clusters k, as well as the method for measuring the distance between observations. The 'optimal' values for these are dependent on the data, and there is no definite way to determine the 'best' clustering.

For data $\mathbb{X} = [\mathbf{X}_1 ... \mathbf{X}_n]$ we fix a distance $\triangle$, the number of clusters k, and a stopping criterion. We then pick k vectors $\bar{\mathbf{X}}_v$ as our cluster centers, and for each $\mathbf{X}_i$, find the $\bar{\mathbf{X}}_v$ it is closest to and assign it to that cluster. Next, for each $v \leq k$ we consider all observations in the cluster around $\bar{\mathbf{X}}_v$, calculate the mean of these observations, and replace $\bar{\mathbf{X}}_v$ by this mean. This process is repeated until the stopping criterion is met.

In this analysis, we have that the data $\mathbb{X} = [\mathbf{X}_1 ... \mathbf{X}_n]$ consists of the n spectra of a patient, where $\mathbf{X}_i$ is the mass spectrum at a single point. The 171,157 m/z values are the variables. As we do not know how many tissue types there may be in each cancer, we perform this initial explorative analysis on a k value of 2, and we use the Euclidean distance as the distance metric.

3.2 Principal component analysis

To determine points of variance that may be determining clusters, we used principal component analysis (PCA). PCA is a dimension that decomposes the data into what are called principal components (PC), which are linear combinations of the original variables. We derive the principal components from the sample covariance matrix S of the data $\mathbb{X}$ through its spectral decomposition. The k^{th} principal component is the projection of the data in the direction of the k^{th} eigenvector of S.

The first principal component direction vector is such that it has maximal variance, and the second will similarly have maximal variance, with the restriction that it must be orthogonal to all previous component direction vectors. In this analysis, we are interested in the variables that have the biggest contribution to the first principal component, as they will have high variability.

4 Cluster patterns of patients individually

Our first aim of this analysis is to determine cluster patterns for patients based on their spectra.

4.1 LNM negative patients

Figure 3 depicts the cluster allocations of each point on the cancer for the 8 LNM negative patients used for this analysis. These patients are patient 157, 158, 159, 160, 161, 162, 163, and 164. This subset is used to illustrate the results of the analysis. We see here that for each patient the data splits into one larger cluster, and one smaller cluster. We also see that the spatial location of the smaller cluster does not appear to be random, that points in the smaller cluster tend to be near each other. Though intuitive otherwise, this implies that the spatial information may be helpful for future classification.

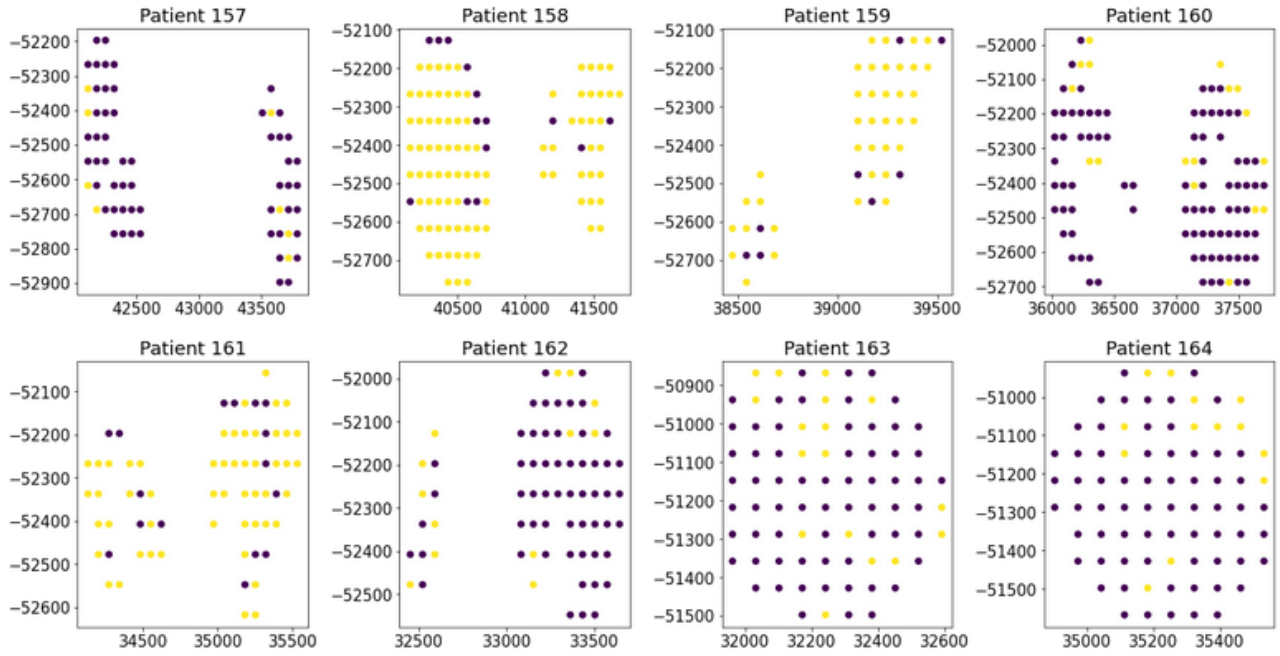


Figure 3: Cluster allocations from k-means clustering with 2 clusters of each point on the cancers for each LNM negative patient. Top L-R: Patient 157, 158, 159, and 160. Bottom L-R: Patient 161, 162, 163, and 164.

4.2 LNM positive patients

Figure 4 depicts the cluster allocations of each point on the cancer for the 8 LNM positive patients used for this analysis. Again, a smaller subset is used to illustrate the results of the analysis, and the patients included are patients 534, 541, 543, 544, 545, 550, 554, and 558. We see here, as with the LNM negative case, that the points of each cancer split into one larger cluster and one smaller cluster. Similarly, we again see spatial grouping of the 2 clusters.

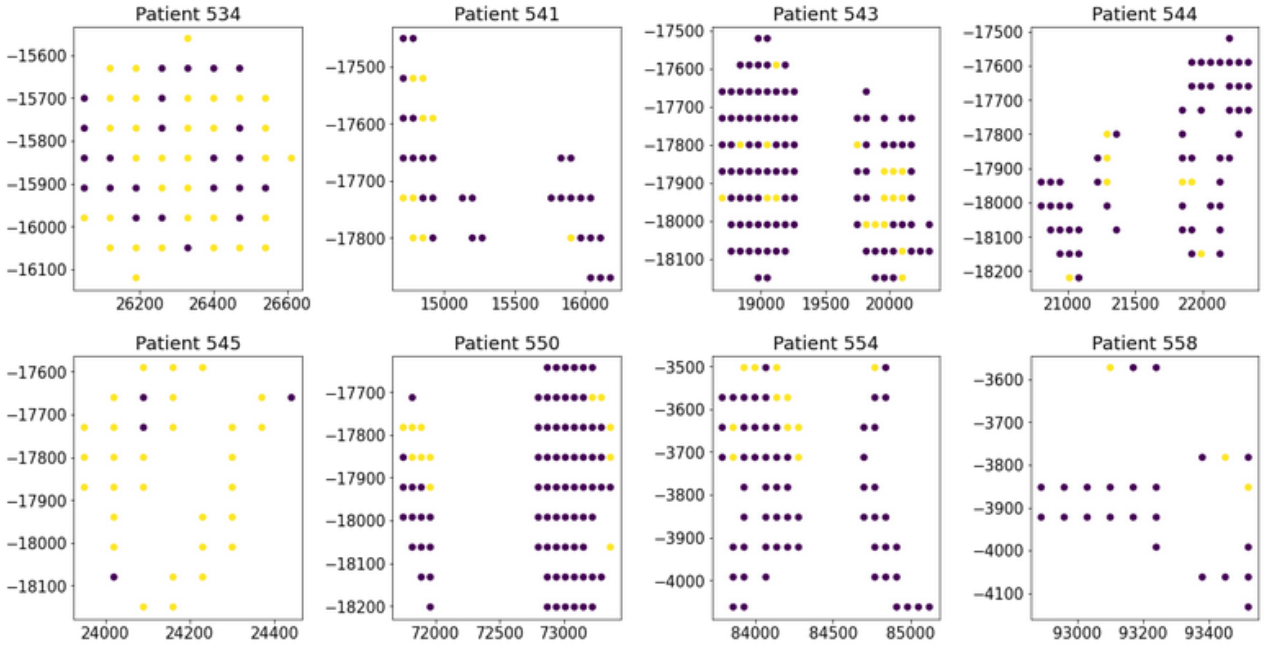


Figure 4: Cluster allocations from k-means clustering with 2 clusters of each point on the cancers for each LNM positive patient. Top L-R: Patient 534, 541, 543, and 544. Bottom L-R: Patient 545, 550, 554, and 558.

4.3 Principal component analysis of patient 158

For the sake of simplicity, we momentarily consider just patient 158. We perform PCA on patient 158's mass spectra, and we obtain the first 10 principal components. Figure 5 shows the contribution to variance of these principal components. We can see that the first principal component contributes to nearly 80% of all total variance, which is very high.

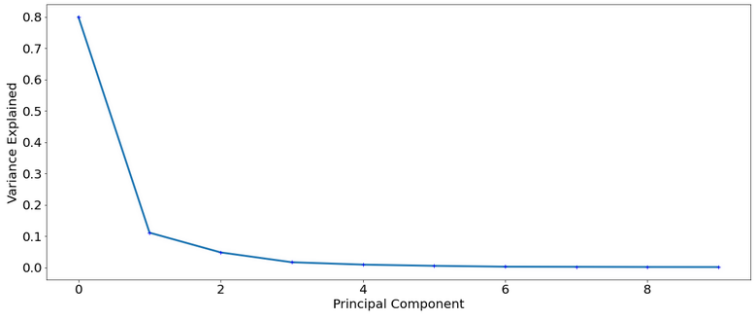


Figure 5: Contribution to variance of the first 10 principal components for patient 158's mass spectra.

To determine if there were only a few or many variables contributing to this high variance, a histogram of the size of the entries of the first direction vector by each variable was plotted. The left plot in Figure 6 shows the histogram of the size of entries from all variables. Due to the majority of variables having a near-zero size, it is impossible to see the variables with large entries. The right plot in Figure 6 shows the size of the entries of these other variables. We can see that we have 2 variables with a size of around 0.35, and another 2 variables with a size of approximately 0.3 that stand apart from the remaining variables. These 4 variables, in order of absolute size, are 842.463, 842.452, 842.474, and 842.440.

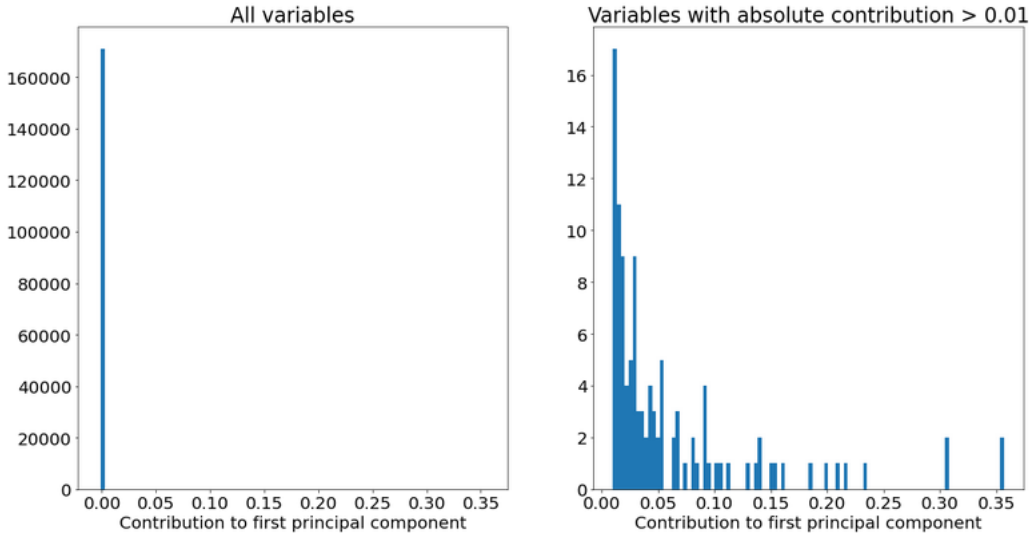


Figure 6: Histogram of the size of entries of the first direction vector. All variables (left) and all variables where absolute size is greater than 0.01 (right).

4.4 Principal components of LNM negative patients

Performing PCA on all LNM negative patients and determining the 4 variables that have the largest entry to the first direction vector, we obtain the results in Table 1. Here we note that the largest entries to the first direction vector is found at the 842 mass for all LNM negative patients.

Patient	Highest 4 contributing variables
157	842.486, 842.474, 842.497, 842.463
158	842.463, 842.452, 842.474, 842.440
159	842.452, 842.463, 842.44, 842.474
160	842.452, 842.463, 842.474, 842.44
161	842.463, 842.474, 842.452, 842.486
162	842.463, 842.474, 842.452, 842.486
163	842.463, 842.474, 842.452, 842.486
164	842.463, 842.474, 842.486, 842.452

Table 1: Variables with the largest entry to the first direction vector (in order of size) for all LNM negative patients.

4.5 Principal components of LNM positive patients

Performing PCA on all LNM positive patients individually and determining the 4 variables that have the largest entry to the first direction vector, we obtain the results in Table 2. Though the largest entries were found at the 842 mass for all LNM negative patients, we do not see the same pattern for the LNM positive patients. We instead have 3 main masses of interest around 842, 944, and 1198, that all appear multiple times. We also have a 4th, less common, mass at 861 that appears only for patient 558.

Patient	Highest 4 contributing variables
534	842.477, 944.51, 842.466, 944.498
541	1198.699, 1198.685, 1198.714, 1198.67
543	944.523, 944.535, 944.51, 944.547
544	1198.699, 1198.685, 1198.714, 1198.729
545	842.477, 944.523, 944.535, 842.488
550	842.466, 842.477, 842.455, 842.488
554	944.498, 944.51, 944.485, 944.523
558	861.068, 861.08, 861.057, 825.098

Table 2: Variables with the largest entry to the first direction vector (in order of size) for all LNM positive patients.

A possible cause of this difference within the LNM positive population could be due to the progression of the metastasised cancer. Further communication with Professor Hoffman is required to determine the cause of this variation.

Though biological reasoning is important for application to patients, it is also interesting to determine the mathematical reasoning for the variation. Figure 7 shows the 842, 944, and 1998 masses for each spectrum for the group of LNM negative patients and group of LNM positive patients.

In Figure 7 each row is plotted on the same scale to demonstrate the difference in variance between masses.

In the group of LNM negative patients, we observe that the mass around 842 has a maximum intensity of nearly 80, whereas the masses around 944 and 1198 have peaks of around 7 and 4, respectively. These plots demonstrate why we see the variables that make up the 842 mass as the largest entries to the first direction vector, as there is a clear large relative intensity.

However, in the group of LNM positive patients, the mass around 842 has a maximum intensity of only 8, and the masses around 944 and 1198 have peaks around 4. We see here that comparatively, the 842 mass may not always have a significantly higher variance than the masses around 944 and 1198.

We do not yet know the peptides that these masses represent, and therefore the biological significance of these findings.

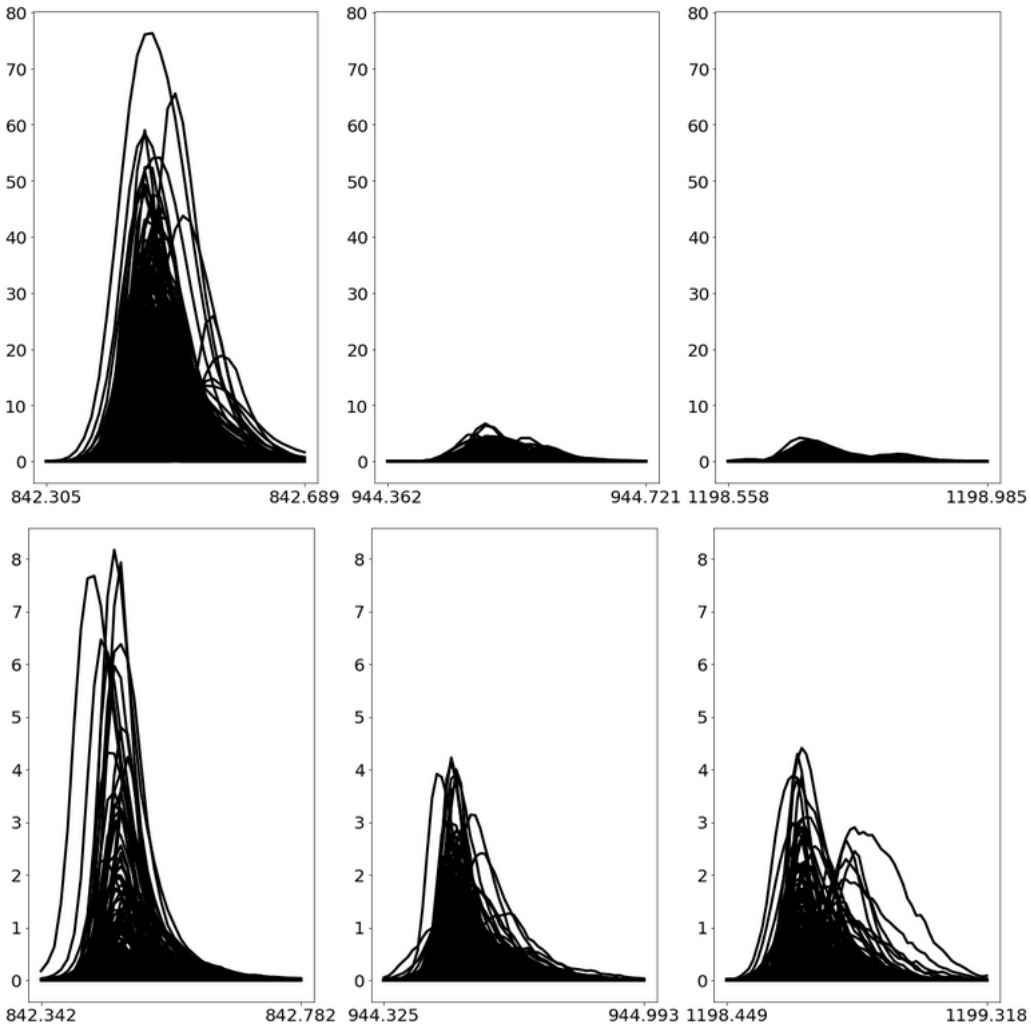


Figure 7: Mass spectra for all LNM negative patients (top) and all LNM positive patients (bottom) at masses 842 (left), 944 (middle), and 1198 (right).

5 Cluster patterns of LNM negative and positive groups

Having looked at the cluster patterns of each patient individually, we next want to determine how these compare to the cluster patterns of the group of LNM negative patients and group of LNM positive patients.

5.1 Cluster patterns and PCA of all LNM negative patients

We again perform k-means clustering on the group of patients, as well as PCA. From PCA we obtain 3 variables that have a relatively large entry to the first direction vector: 842.463, 842.452, and 842.474, in order of size. It is not surprising, and was even expected, that the top contributors to the variance are variables from the 842 mass, given this was seen for all LNM negative patients when analysed individually.

Similarly, we see that the cluster allocations for an individual mass spectrum typically stays the same. If a mass spectrum was in the smaller cluster when clustered with the remaining mass spectra from its respective patient, it is more than likely in the smaller cluster when clustered with the mass spectrum from all patients. We can see this in Figure 8, where all but a couple of mass spectra from patient 158 stay in the same cluster allocations, whether the clustering is done as an individual patient or as part of the group of all LNM negative patients. The mass spectra that do change clusters are those where the maximum intensity of the mass is close to the minimum intensity of the cluster with higher intensities and also close to the maximum intensity of the cluster with lower intensities. That is to say, these mass spectra fall between clusters, and hence the change in cluster allocation is not surprising.

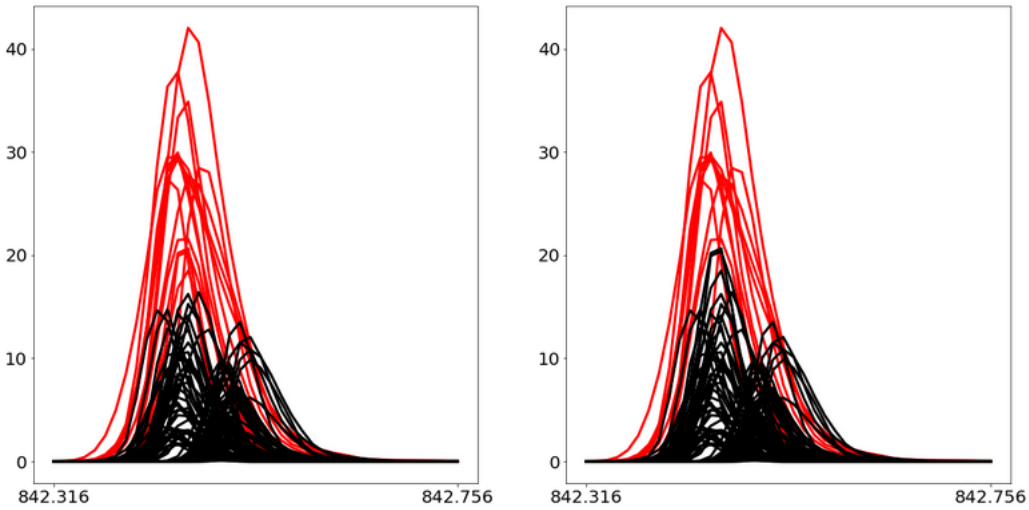


Figure 8: Mass spectrum of patient 158 at the 842 mass. Coloured by cluster allocations. Clustering done as an individual patient (left) and as a group of all LNM negative patients (right).

We see this adherence to population for all LNM negative patients and can thus conclude that there is little variability within the group of LNM negative patients.

5.2 Cluster patterns and PCA of all LNM positive patients

When we perform this subsequent analysis on the group of LNM positive patients, we again see variance within the population. Performing PCA, we see 2 variables with an entry to the first vector direction of over 0.3, and another 2 variables with a size of entry of over 0.25. These variables are 842.466, 842.477, 842.455, and 842.488, in order of size. This is not altogether surprising, as it is shown in Figure 7 the intensity of the 842 mass is nearly double that of the 944 and 1998 masses. It is still interesting however, as it doesn't represent the differences we saw when clustering at the individual level.

The difference within the population of LNM positive patients can also be seen when comparing cluster allocations of individual mass spectra when clustered for an individual patient against being clustered with all LNM positive patients. Figure 9 shows the cluster allocations for the mass spectra of patient 534 around the 842 mass between clustering at the patient level against clustering with all LNM positive patients. We can see that the mass spectra that have the highest intensity stay in the red coloured cluster, however, most of the other mass spectra initially in the red coloured cluster change to be in the black coloured cluster. This is particularly interesting as the variables with the largest entries to the first vector direction at the individual level for patient 534 were from this mass around 842.

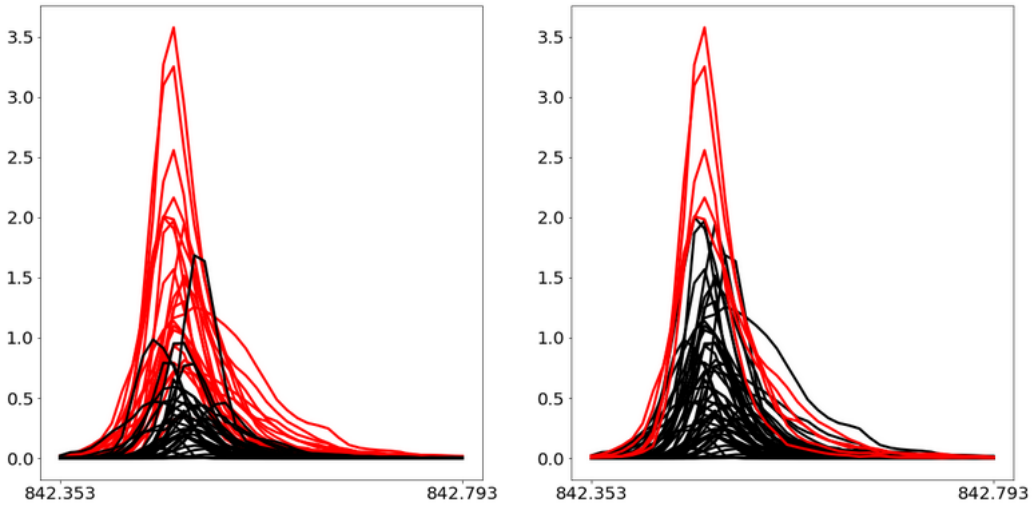


Figure 9: Mass spectrum of patient 534 at the 842 mass. Coloured by cluster allocations. Clustering done as an individual patient (left) and as a group of all LNM positive patients (right).

6 Discussion and Conclusion

This analysis has identified 3 potential peptides of interest for future classification, corresponding to the masses around 842, 944 and 1198 m/z. Through comparison of clustering at the individual level and clustering of the population and PCA we have also identified significant variance within the LNM positive population, and cohesiveness within the LNM negative population.

While finalising this report, we learnt that the 842 mass is a trypsin peptide, the 944 mass is a histone

peptide, and the 1198 mass is a actin peptide. Trypsin is the peptide applied to the tissue sample as part of the MALDI-MSI process. Though the stark difference in variance between the LNM positive and negative populations is interesting, the peptide itself has no biological significance. The significance of the histone and actin peptides are still unknown.

The next step for this project will be to remove an interval around the mass at $842 m/z$ and complete this analysis again. Further experimental study will be required to determine what interval to leave out. We will also look at binning the data around these masses and taking an average across the bin. This will not only reduce the dimensionality of the data, but also capture the entirety of a mass, rather than a point along it.

References

Jacques, SM et al. (Jan. 1998). "Interinstitutional surgical pathology review in gynecologic oncology: I. Cancer in endometrial curettings and biopsies". In: *International journal of gynecological pathology : official journal of the International Society of Gynecological Pathologists* 17.1, pp. 36–41. ISSN: 0277-1691. DOI: 10.1097/00004347-199801000-00007. URL: https://doi.org/10.1097/00004347-199801000-00007.

Mittal, Parul et al. (2016). "Lymph node metastasis of primary endometrial cancers: Associated proteins revealed by MALDI imaging". eng. In: *Proteomics (Weinheim)* 16.11-12, pp. 1793–1801. ISSN: 1615-9853.

Rungruang, Bunja and Alexander B Olawaiye (2012). "Comprehensive surgical staging for endometrial cancer". eng. In: *Reviews in obstetrics and gynecology* 5.1, pp. 28–34. ISSN: 1941-2797.

Todo, Yukiharu et al. (2010). "Risk factors for postoperative lower-extremity lymphedema in endometrial cancer survivors who had treatment including lymphadenectomy". eng. In: *Gynecologic oncology* 119.1, pp. 60–64. ISSN: 0090-8258.

Winderbaum, Lyron et al. (2016). "Classification of MALDI-MS imaging data of tissue microarrays using canonical correlation analysis-based variable selection". eng. In: *Proteomics (Weinheim)* 16.11-12, pp. 1731–1735. ISSN: 1615-9853.