Technical Brief
Classification of MALDI-MS imaging data of tissue microarrays using Canonical Correlation Analysis based variable selection

Lyron Winderbaum,* Inge Koch, Parul Mittal, and Peter Hoffmann

The University of Adelaide

(Dated: March 4, 2016)

Abstract

Applying MALDI-MS imaging to tissue microarrays (TMAs) provides access to proteomics data from large cohorts of patients in a cost- and time-efficient way, and opens the potential for applying this technology in clinical diagnosis. The complexity of these TMA data – high-dimensional low sample size (HDLSS) – provides challenges for the statistical analysis, as classical methods typically require a non-singular covariance matrix which cannot be satisfied if the dimension is greater than the sample size. We use TMAs to collect data from endometrial primary carcinomas from 43 patients. Each patient has a lymph node metastasis (LNM) status of positive or negative, which we predict on the basis of the MALDI-MS imaging TMA data. We propose a variable selection approach based on Canonical Correlation Analysis (CCA) which explicitly uses the LNM information. We apply LDA to the selected variables only. Our method misclassifies 2.3–20.9% of patients by leave-one-out (LOO) cross-validation and strongly outperforms LDA after reduction of the original data with principle component analysis (PCA).

Keywords: Canonical Correlation Analysis; Classification; Endometrial cancer; MALDI-MS imaging; Variable ranking

[*] **E-mail:** lyron.winderbaum@student.adelaide.edu.au;

Correspondence: Lyron Winderbaum, Adelaide Proteomics Centre, North Terrace Campus, Level 1 Molecular Life Sciences, The University of Adelaide, SA 5005 Australia

Fax: +61 (0)8 8313 4362

Abbreviations: CCA, Canonical Correlation Analysis; **TMA,** Tissue Microarray; **LDA,** Linear Discriminant Analysis; **LNM,** Lymph Node Metastasis; **FFPE,** Formalin-Fixed, Paraffin-Embedded; **HDLSS,** High-Dimension Low Sample Size

MALDI-MS imaging enables the relative quantification and spatial expression profiling of thousands of peptides within and between tissues. Using tissue microarrays (TMAs) can facilitate the acquisition of data from large cohorts of patients [1], but relative to the number of resolvable ion species, the sample sizes remain small – less than 100 patients represented in a typical study, although there are some notable exceptions [2]. Classification at a patient level of such data is of clinical interest, but classical approaches to classification, such as linear discriminant analysis (LDA), do not apply without modifications such as prior feature selection or regularisation. In this paper we suggest a method for dimension reduction via variable selection prior to classification and demonstrate its effectiveness when applied to the data of [3], which consist of MALDI-MS imaging data collected from TMAs of formalin-fixed paraffin-embedded (FFPE) primary endometrial carcinoma tissue. For these endometrial cancer data it is of interest to distinguish between patients with lymph node metastasis (LNM) and those without.

Endometrial cancer is the most common gynaecological malignancy in Australia with 2256 diagnosed cases in 2010 and 381 associated deaths in 2011 [4]. The presence or absence of lymph node metastasis (LNM) is the most important prognostic factor in endometrial cancer as patients with localised disease have a 5 year survival rate of 96%, which drops to just 17% for patients with metastatic disease [5]. Accurately staging endometrial cancer is difficult and a large percentage of patients are misclassified prior to treatment [6]. Although the presence of LNM is confirmed in only around 15% of cases [7, 8], the majority of endometrial cancer patients undergo radical treatment including the removal of pelvic lymph nodes as a precautionary measure to compensate for our current inability to accurately stage the disease. Lymph node removal is associated with significant complications including lower extremity lymphoedema, which has been described in up to 38% of patients [9]. A classification system based around predictive tissue markers of metastasis would greatly benefit stage I endometrial cancer patients by helping determine optimal treatment strategies that avoid unnecessary, invasive procedures.

The sample preparation of the endometrial TMAs included citric acid antigen retrieval [10], trypsin digestion, addition of internal calibrants [11], and matrix deposition. Further details on the data acquisition for the endometrial data are available in [3]. As part of the data acquisition peak-picking was performed through proprietary software (flexControl v3.0.1 and flexImaging v4.0.1, Bruker Daltonik, `http://www.bruker.com`). All our analyses are carried out solely on the monoisotopic peaks detected by this method. These peak-data are about two orders of magnitude smaller and result in much faster computation times while preserving important information.

Two sections (technical replicates) of two TMAs were analysed and tumour regions annotated by a pathologist. In this analysis, the only information we use about each spectrum is which patient it belongs to and whether the pathologist's annotations include it as tumour. After consideration of the patient clinical data it was determined that 43 patients suitable for the study are represented across the two TMAs. Of these 43 patients, 16 are LNM positive, 27 are negative. Details on the endometrial cancer project from which these data originate are available in [3].

We construct 0.25Da wide, non-overlapping bins which are the variables in our analyses. In addition we repeat the analyses for shifted bins. Our final prediction combines the corresponding results from the individual analyses, using a majority rule. Details regarding the choice of bins, shifted-bin analyses and majority rule are given in the supplementary information.

For each patient, spectra were averaged for each m/z bin. These averages are assembled into data matrices with 43 columns corresponding to the patients represented in the study, and 4582 rows corresponding to m/z bins with at least one peak in them in at least one spectrum (4570 and 4584 in the two shifted-bin analyses respectively). This resulting data matrix is what will be used in all following analyses. We centre the rows of these data matrices, and refer to these centred data as X.

As previously mentioned, classical approaches to classification such as LDA do not apply to HDLSS data without modifications. Different solution paths exist in the literature which include replacing the classical LDA by its Naive Bayes version, see [12, Section 13.2.1]. Other methods that are directly applicable to HDLSS data include Distance Weighted Discrimination of [13], kernel methods – see [14], and regularisation approaches which make use of generalised inverses, ridge-like estimates of the covariance matrix, or Lasso and elastic net type classification – see [15, 16] and references therein.

A different solution path consists of reducing the number of variables or features first, and then applying the chosen classifier to the reduced data. In MALDI-MS imaging data PCA is a popular feature reduction method, see [17, 18] and references therein. More recently [19] have suggested the use of PCA on MALDI-MS imaging data from TMAs in a classification context. For a generic reference to PCA see [20] and [12, Chapter 2 and Section 4.8].

Other feature reduction approaches used in proteomics include the application of univariate tests such as Wilcoxon rank-sum statistics or t-tests – see [21, 22]. For high-dimensional data [23] ranked the variables by applying t-tests simultaneously to all variables and analysed the performance of this approach. We follow a related path, based on Canonical Correlation Analysis (CCA), and select the 'best' original variables in a supervised manner prior to classification. The selected variables do not require any interpretation, as they are simply the m/z bins. Our approach contains the approach of [23] as a special case, and does not require tuning parameters as regularisation-type methods do.

In CCA the data are split into two sets of variables and the linear combinations of variables in each subset with the strongest correlation are determined. Typically this correlation is much stronger than that of any individual pair of variables from the two subsets. In a regression context the correlation between a response and the predictor variables shows the relative strength of the response on the individual predictors. In a classification framework, we use class labels of observations as one subset of variables, and all other variables as the second subset. For our data, the class labels are the LNM status, and the variables are all m/z bins.

It is beyond the scope of this paper to review the different approaches mentioned above or to show detailed comparisons. However, because of the repeated use of PCA for MALDI-MS data, we include a performance comparison of our proposed method with that based on feature reduction with PCA. We note that our CCA-based variable selection is supervised, while PCA does not use the class labels in its feature reduction. As a consequence PCA may not find features that are optimal for predicting class membership. In addition, our suggested CCA-based approach simply ranks the original variables, here m/z bins, and picks the best, while PCA transforms the original variables and results in features that are linear combinations of all m/z bins. While we expect to relate the 'best' CCA masses to biomarkers, no such interpretation is possible for the PCA features. In our results, shown in Figure 1 and Figure 2, we will see the dramatic improvement in prediction that can be achieved with the supervised CCA-based variable selection compared to the unsupervised PCA-based dimension reduction.

In the context of MALDI-MS data, CCA has been used in finding correlations between datasets collected from the same tissue using different imaging techniques [24]. These authors use CCA in its standard form: finding correlations between two subsets of variables that belong to a natural partition – m/z values collected using one imaging technique, and m/z values collected using the second imaging technique. CCA has also been used in a classification context [25], although in an exploratory manner – plotting the strongest correlation between their data and their prediction variable to observe a linear relationship.

In this paper we also use the first and strongest correlation obtained from CCA, but we deviate from the conventional approach in that we explicitly use the information inherent in the first direction vector for variable selection. CCA-based variable selection as suggested by [26] is further described in [12, Sections 13.3.1 and 13.3.3], and references therein. We modify their approach for our data, essentially by using centred labels. Here we briefly introduce our variable ranking approach, but further details, motivation, and comments can be found in the supporting information. Let X be a $d \times n$ centred data matrix whose columns represent observations (patients), and whose rows represent variables (m/z bins). Let $\boldsymbol{y}$ be a $1 \times n$ vector of (centred) class labels $\frac{-2n_+}{n}$ and $\frac{2n_-}{n}$ corresponding to the same n observations as the columns of X, where n_+ (n_-) is the number of observations in the positive (negative) class. The key idea is to exploit the relationship between X and $\boldsymbol{y}$. Details describing how we use CCA are included in the supporting information, but essentially we will be interested in the vector $\boldsymbol{\phi}$, which satisfies

$$XX^T\boldsymbol{\phi} = X\boldsymbol{y}^T. \tag{1}$$

The vector $\boldsymbol{\phi}$ contains the weights for the linear combination of variables, or rows of X, which gives rise to the strongest absolute correlation with $\boldsymbol{y}$. Variable reduction to k variables is obtained by selecting variables corresponding to the k highest entries of $\boldsymbol{\phi}$ by absolute value.

In what follows, LDA refers to Fisher's linear classification as described in [12, Section 4.3]. LDA finds the direction which maximises the between-class variance and simultaneously minimises the within-class variance of the data. We will use the term CCA-LDA to refer to the process of first using the CCA variable selection to perform a dimension reduction step, and then apply LDA to the selected variables. Similarly, we will use the term PCA-LDA to refer to the process of first performing PCA dimension reduction, and then apply LDA to the selected variables. We measure the performance of these methods by Leave-One-Out (LOO) cross-validation. LOO cross-validation is an iterative process where in each iteration one observation is left out, the classification rule is trained on the remaining data, and tested on the left-out observation – assigning it a class label. This process is repeated until all observations have been assigned class labels in the 'testing' step. Incorrectly assigned observations are counted as LOO misclassification. We use LOO cross-validation rather than a more traditional training/ testing design (where separate datasets are used for the training and testing steps) because of the relatively small number of observations ($n = 43$).

We examined the effect of the parameters such as *regions of tissue* and *data form* as variations of the computational pipeline. In particular we consider the variations listed below and implement all possible combinations of these choices in order to determine which parameters are most important for classification.

- Data form: using raw intensities (I), log-intensities ($\log(I+1)$), peak areas, signal-to-noise ratios or binary presence/ absence of peaks.
  - For non-binary data forms, absence of peaks: When averaging spectra in which some m/z bins contains no peaks, the absence of peaks is included as a zero value, or ignored and not be included in the average.
  - For binary data, spatial smoothing: The spatially smoothed binary data as described in [27] is used with smoothing parameters 0, 0.15, or 0.25 where 0 corresponds to no smoothing.
- Regions of tissue: all spectra, or spectra from annotated tumour regions only. Restricting to only annotated tumour regions represents a reduction in the total number of spectra from 123561 to 45877 ($\sim 37\%$) in the endometrial TMA data.

Our analyses showed that *data form* was the only factor that consistently affected classification performance, with log-intensity data performing best, and binary data typically performing second best.

Overall, the CCA-LDA outperformed PCA-LDA. A typical case can be seen in Figure 1 in which we compare our CCA-LDA with PCA-LDA using the same parameter settings – log-intensity data with zeroes included for absent peaks in the averaging step, and restricting to only annotated tumour regions. Figure 1 shows the LOO misclassification (out of 43 patients) of CCA-LDA and PCA-LDA on the y-axis against the number of variables or dimensions on the x-axis (used in the LDA step). In the case shown in Figure 1, CCA-LDA strictly outperforms PCA-LDA and it can be seen that CCA-LDA achieves its most parsimonious minimum misclassification of 5 by selecting 19 variables. Figure 2 shows the most parsimonious minimum misclassifications for each of the 22 combinations listed in the dot points above, when using PCA-LDA and CCA-LDA. The minimum misclassifications of CCA-LDA shown in Figure 2 vary between 1 (2.3%) and 9 (20.9%), while the minimum misclassifications of PCA-LDA vary between 7 (16.3%) and 18 (41.9%).

The improved performance of our CCA-based variable selection over the standard PCA dimension reduction, which uses linear combinations of variables, is likely due to the fact that PCA is a purely variance-based method, which ignores any information about the clinical outcome of interest (LNM). It is possible that the high-variance components of the data, which PCA selects, may not include the crucial classification information necessary in order to predict the clinical outcome accurately. CCA variable selection in contrast explicitly takes into account the clinical outcome when selecting the variables, and this could account for its improved performance in comparison to PCA.

Another advantage of CCA variable selection over that with PCA is the interpretability of the variable-reduced data – the PCA variable-reduced data is produced by projection and so the new variables are linear combinations of the original variables, making them difficult to interpret. The variables selected via our CCA-based approach are a subset of the original variables and so preserve their interpretation as m/z bins. This interpretation is of use in designing follow-up biomarker studies. Some of the specific m/z bins that are highly ranked in the CCA-based variable selection are discussed in [3].

In conclusion, we have presented a variable selection method for MALDI-MS imaging data on TMAs, and demonstrated that this variable selection method combined with LDA achieves a LOO misclassification of between 2.3% and 20.9% for lymph node metastasis status in the endometrial cancer data of [3].

This research was carried out in collaboration with work supported by the Ovarian Cancer Research Foundation. The authors would also like to thank Professor Martin Oehler for access to the samples, and Associate Professor Gurjeet Kaur for her help with histopathological annotations.

[1] Rita Casadonte and Richard M Caprioli. Proteomic analysis of formalin-fixed paraffin-embedded tissue by MALDI imaging mass spectrometry. *Nature protocols*, 6(11):1695–1709,

2011.

[2] Stefan Steurer, Carina Borkowski, Sinje Odinga, Malte Buchholz, Christina Koop, Hartwig Huland, Michael Becker, Matthias Witt, Dennis Trede, Maryam Omidi, et al. MALDI mass spectrometric imaging based identification of clinically relevant signals in prostate cancer using large-scale tissue microarrays. *International Journal of Cancer*, 133(4):920–928, 2013.

[3] Parul Mittal, Manuela Klingler-Hoffmann, Georgia Arentz, Lyron Winderbaum, Noor A. Lokman, Chao Zhang, Gurjeet Kaur, Martin Oehler, and Peter Hoffmann. MALDI imaging of primary endometrial cancers reveals proteins associated with lymph node metastasis. Submitted to Proteomics: Imaging Mass Spectrometry Special Issue.

[4] Gynaecological cancers in australia: an overview, 2012. Canberra: AIHW.

[5] Bunja Rungruang and Alexander B Olawaiye. Comprehensive surgical staging for endometrial cancer. *Reviews in obstetrics & gynecology*, 5(1):28–34, 2012.

[6] Suzanne M Jacques, Faisal Qureshi, Adnan Munkarah, and W Dwayne Lawrence. Interinstitutional surgical pathology review in gynecologic oncology I. cancer in endometrial curettings and biopsies. *International journal of gynecological pathology*, 17(1):36–41, 1998.

[7] C Paul Morrow, Brian N Bundy, Robert J Kurman, William T Creasman, Paul Heller, Howard D Homesley, and James E Graham. Relationship between surgical-pathological risk factors and outcome in clinical stage I and II carcinoma of the endometrium: a gynecologic oncology group study. *Gynecologic oncology*, 40(1):55–65, 1991.

[8] WT Creasman, F Odicino, P Maisonneuve, MA Quinn, U Beller, JL Benedet, APM Heintz, HYS Ngan, and S Pecorelli. Carcinoma of the corpus uteri. *International Journal of Gynecology & Obstetrics*, 95:S105–S143, 2006.

[9] Yukiharu Todo, Ritsu Yamamoto, Shinichiro Minobe, Yoshihiro Suzuki, Umazume Takeshi, Makiko Nakatani, Yukiko Aoyagi, Yoko Ohba, Kazuhira Okamoto, and Hidenori Kato. Risk factors for postoperative lower-extremity lymphedema in endometrial cancer survivors who had treatment including lymphadenectomy. *Gynecologic oncology*, 119(1):60–64, 2010.

[10] Johan O. R. Gustafsson, Martin K. Oehler, Shaun R. McColl, and Peter Hoffmann. Citric acid antigen retrieval (CAAR) for tryptic peptide imaging directly on archived formalin-fixed paraffin-embedded tissue. *Journal of Proteome Research*, 9(9):4315–4328, July 2010.

[11] Johan O.R. Gustafsson, James S. Eddes, Stephan Meding, Tomas Koudelka, Martin K. Oehler, Shaun R. McColl, and Peter Hoffmann. Internal calibrants allow high accuracy peptide match-

ing between MALDI imaging MS and LC-MS/MS. *Journal of Proteomics*, 75(16):5093 – 5105, 2012. Special Issue: Imaging Mass Spectrometry: A Users Guide to a New Technique for Biological and Biomedical Research.

[12] Inge Koch. *Analysis of Multivariate and High-Dimensional Data*, volume 32 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2013.

[13] J.S. Marron, M.J. Todd, and J. Ahn. Distance-weighted discrimination. *American Statistical Association*, 102(480):1267–1271, 2007.

[14] Sebastian Mika, Gunnar Rätsch, Jason Weston, Bernhard Schölkopf, and Klaus-Robert Müller. Fisher discriminant analysis with kernels. *Neural networks for signal processing IX*, 1(1):1, 1999.

[15] Ping Xu, Guy N. Brock, and Rudolph S. Parrish. Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. *Computational Statistics & Data Analysis*, 53(5):1674–1687, 2009.

[16] Martin Vincent and Niels Richard Hansen. Sparse group lasso and high dimensional multinomial classification. *Computational Statistics & Data Analysis*, 71:771–786, 2014.

[17] Gregor McCombie, Dieter Staab, Markus Stoeckli, and Richard Knochenmuss. Spatial and spectral correlations in MALDI mass spectrometry images by clustering and multivariate analysis. *Analytical chemistry*, 77(19):6118–6124, 2005.

[18] Soren-Oliver Deininger, Matthias P. Ebert, Arne Futterer, Marc Gerhard, and Christoph Rocken. MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. *Journal of Proteome Research*, 7(12):5230–5236, 2008. PMID: 19367705.

[19] Nadine E Mascini, Gert B Eijkel, Petra ter Brugge, Jos Jonkers, Jelle Wesseling, and Ron MA Heeren. The use of mass spectrometry imaging to predict treatment response of patient-derived xenograft models of triple-negative breast cancer. *Journal of proteome research*, 14(2):1069–1075, 2015.

[20] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[21] Sandra Rauser, Claudio Marquardt, Benjamin Balluff, Sören-Oliver Deininger, Christian Albers, Eckhard Belau, Ralf Hartmer, Detlev Suckau, Katja Specht, Matthias Philip Ebert, Manfred Schmitt, Heinz Höfler Aubele, and Axel Walch. Classification of HER2 receptor status in breast cancer tissues by MALDI imaging mass spectrometry. *Journal of proteome*

research, 9(4):1854–1863, 2010.

[22] M. Reid Groseclose, Pierre P. Massion, Pierre Chaurand, and Richard M. Caprioli. High-throughput proteomic analysis of formalin-fixed paraffin-embedded tissue microarrays using MALDI imaging mass spectrometry. *Proteomics*, 8(18):3715–3724, 2008.

[23] Jianqing Fan and Yingying Fan. High-dimensional classification using features annealed independence rules. *Annals of Statistics*, 36:2605–2637, 2008.

[24] GB Eijkel, B Kükrer Kaletaş, IM Van Der Wiel, JM Kros, TM Luider, and RMA Heeren. Correlating MALDI and SIMS imaging mass spectrometric datasets of biological tissue surfaces. *Surface and Interface Analysis*, 41(8):675–685, 2009.

[25] Nicoletta Nicolaou, Yun Xu, and Royston Goodacre. Detection and quantification of bacterial spoilage in milk and pork meat using MALDI-TOF-MS and multivariate analysis. *Analytical chemistry*, 84(14):5951–5958, 2012.

[26] Inge Koch and Kanta Naito. Prediction of multivariate responses with a selected number of principal components. *Computational Statistics & Data Analysis*, 54(7):1791–1807, 2010.

[27] Lyron J Winderbaum, Inge Koch, Ove JR Gustafsson, Stephan Meding, Peter Hoffmann, et al. Feature extraction for proteomics imaging mass spectrometry data. *The Annals of Applied Statistics*, 9(4):1973–1996, 2015.

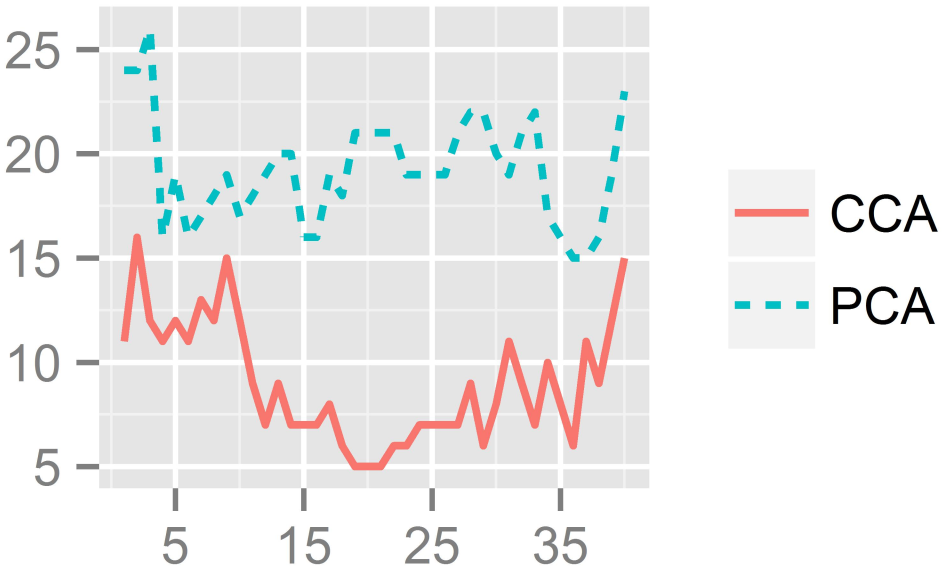

FIG. 1: LOO-misclassification for log-intensity data, including zeroes for absent peaks when averaging and restricting to spectra from tumour regions. The x-axis shows the number of variables (m/z bins) 1 to 40 included in the CCA data, or the number of principle components (linear combinations of m/z bins), 1 to 40 included in the PCA data. For each dimension-reduced data, LDA is performed and the LOO-misclassification shown on the y-axis separately for CCA and PCA. The smallest misclassification of 5 for CCA-LDA occurs when the best 19 variables are selected. The smallest misclassification of 15 for PCA-LDA occurs when the PCA-dimension is 36. A comparison shows that CCA-LDA results in lower misclassification and a more parsimonious model in this case.

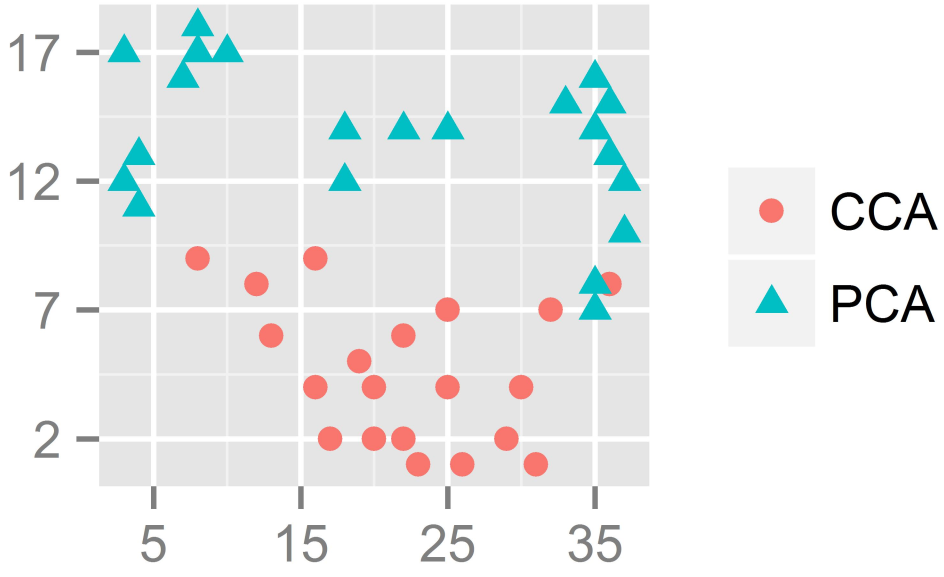

FIG. 2: Best LOO-misclassification for each of the 22 combinations of data type, inclusion/ exclusion of empty values (for non-binary data types), smoothing (for binary data types), and restriction to spectra from tumour annotated tissue. The x-axis shows the minimum number of variables for the CCA data, or the number of principle components of the PCA data, that results in the smallest LOO-misclassification in each scenario. The y-axis shows the number of misclassified patients for that scenario, with CCA in circles and PCA in triangles. For example the circle at $(x, y) = (23, 1)$ indicates that CCA-LDA results in a LOO misclassification of one when the best 23 CCA-selected variables are used in LDA for this particular scenario (binary data with a 0.25-smooth and restricting to spectra from regions annotated as tumour only). The figure clearly shows that CCA-LDA consistently achieves better LOO misclassification than PCA-LDA on these data.