Cork Institute of Technology

A study of stability of statistical analysis of mass spectrometry data

by

Jean-Michel Rubillon - R00107142
Department of Mathematics

Supervised by Dr David Hawe

Submitted in partial fulfilment of the requirements of the

Higher Diploma in Science in Data Science and Analytics

February 2018

Attestation of authorship

This project report is the sole work of Jean-Michel Rubillon unless otherwise indicated and is submitted in the partial fulfilment of the degree of Higher Diploma in Science in Data Science and Analytics. I understand that significant plagiarism, as determined by the examiner, may result in the award of zero marks for the entire assignment. Anything taken from or based upon the work of others has its source clearly and explicitly cited.

Date: 20 February 2018

Signature: Jean-Michel Rubillon

Contents

List of Figures

Chapter 1

Introduction

The determination of the relative molecule quantities within a pharmaceutical product are an essential part of their design and production consistency. This determination is made using mass spectroscopy of samples taken from either a laboratory batch or production. The high cost of running these tests and the delay in getting results means that relatively few samples are analysed at a time. So, on top of the identification issues highlighted in [1] one has to wonder whether the number of samples taken are sufficient to represent the true picture of the product's contents.

Because of the large number of spectral lines available, determining the identity of compounds can be difficult. Typically, this is done using partial least squares (PLS) to identify the most important predictors for identification purposes.

As [2] points out, proteomics, like all high-throughput technologies, is extremely dependent on the ability to quickly and reliably analyze large amounts of experimental data. In the absence of robust statistical and computational methods, proteomic datasets contain significant numbers of false positives, and statements referring to computational analysis of MS/MS data as e.g. "the Achilles heels of proteomics" are common in the literature.

The question therefore arises as to whether the partial least squares (PLS) method is robust enough considering the low sample count and thus the potential for low statistical power of any analysis performed.

The aim of this project is to determine whether PLS is sufficiently robust considering the high variability of mass spectrometry data for the same product being investigated. We will establish the amount of noise that can be tolerated by the method before erroneous predictions occur. We will then examine whether other methods such as principal component

analysis (PCA) could be applied with better noise immunity.

Chapter 2

Literature review

In this chapter we will summarise the state of the art in terms of statistical analysis in the field of proteomics. Proteomics being the study of proteomes which are a sets of proteins produced by an organism [3].

2.1 Mass Spectrometry

Mass spectrometry is an analytical technique enabling the quantification of known materials as well as the identification of unknown compounds within a sample. It also helps in the identification of molecular structure and associated chemical properties.

At its simplest the mass spectrometer generates multiple ions from the sample under test, separates them according to their mass to charge ratio (m/z) and records the relative abundance of each ion type [4] as illustrated in Figure 2.1.1.

Mass spectrometry can be carried out using a variety of different techniques. For illustration we will present one here: **electron ionisation**. This is where gas phase molecules enter the ion source where they are bombarded with free electrons emitted from a filament (Figure 2.1.2). The electrons bombard the molecules causing a hard ionization that fragments the molecule and turn into positively charged particles called ions. The filter continuously scans through the range of masses as the stream of ions come from the ion source. A detector counts the number of ions with a specific mass which in turn generates the mass spectrum [5].

The resultant data analysis is then that of matching the spectra to candidate peptides and inferring the original proteins from these peptides [6]. As [6] points out, one of the major

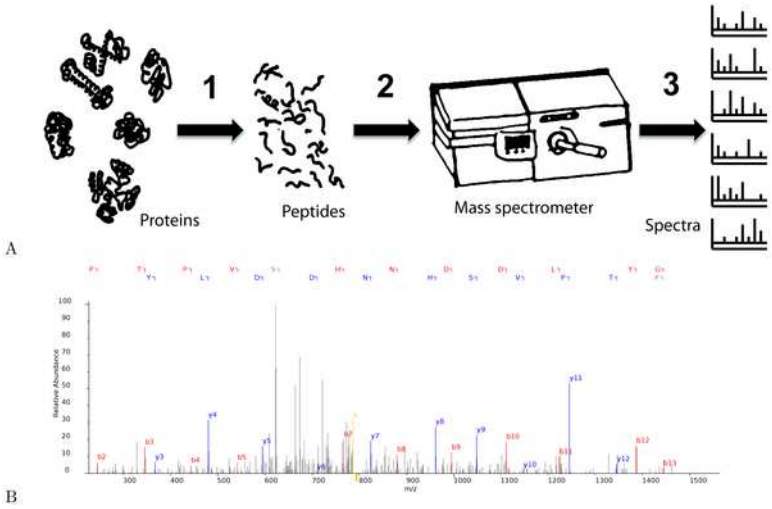

Figure 2.1.1: A simplified graphical representation of the mass spectrometry process [4]

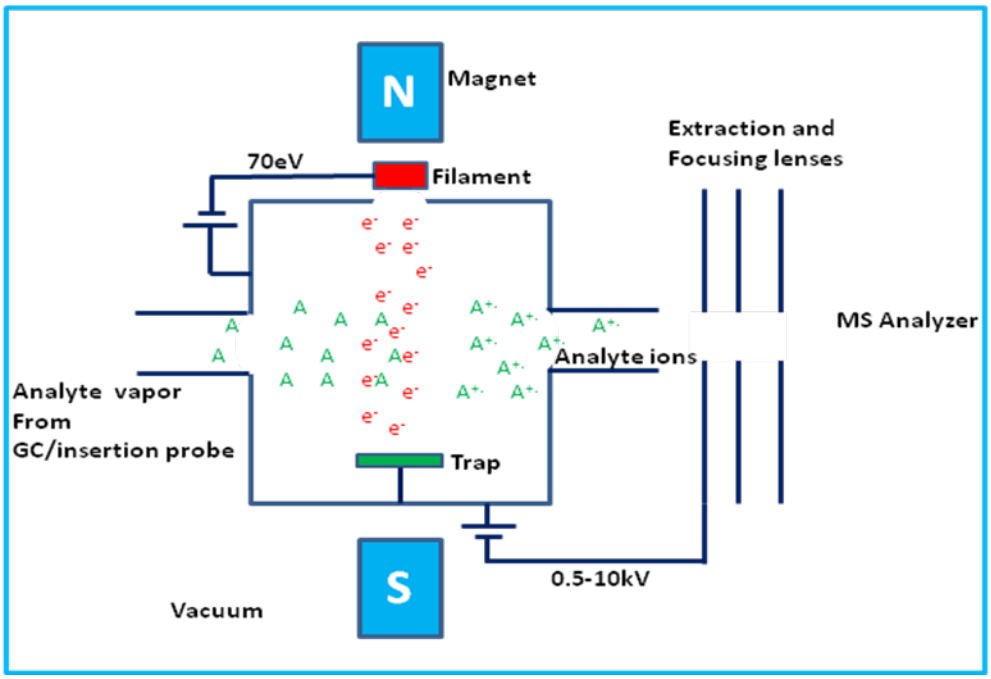

Figure 2.1.2: Schematic diagram of electro ionisation source [5]

challenges in this procedure is the high noise content in the spectra i.e., the spectra have a low signal to noise ratio.

The main issue in proteomics data analysis is the "curse of dimensionality" where $n \ll p$. The number of samples n is much smaller than the number of variables (peptides, proteins, molecules) p.

The properties of proteomic datasets can hamper the development of robust classifiers. It is thus a common approach to reduce the dimensionality of the dataset prior to a multivariate analysis [7]. Various techniques have been used to analyse proteomic data such as support vector machines (SVM), artificial neural networks (ANN), and random forests as partial least squares-discriminant analysis (PLS-DA) and principal components regression-linear discriminant analysis (PCR-LDA) [8].

[8], reminds us that PCA and PLS are dimension reduction methods and, although they provide class separation on a qualitative level, they are not strictly speaking classification methods. They are typically used in conjunction with existing classification methods.

2.2 Principal Component Analysis (PCA)

Dealing with a large amount of variables, it is difficult to identify any interdependence between them. One of the most commonly used multivariate statistical analysis method is Principal Component Analysis (PCA). The principle of PCA is create a minimum set of new variables that describe the variation of the original data. This is achieved by using combinations of linear equations of the original data [9].

PCA is effectively a method for finding the variables that best differentiate data points i.e. the dimensions along which the data points are the most spread-out [10].

In order to choose how many principal components to use, a criterium based on the total captured variance can be used.

In mathematical terms [13], supposing r variables and N samples, the dataset can be represented by:

$$\begin{array}{c} x_{11}\vec{e_1} + x_{12}\vec{e_2} + \cdots + x_{1r}\vec{e_r} \\ x_{21}\vec{e_1} + x_{22}\vec{e_2} + \cdots + x_{2r}\vec{e_r} \\ \vdots \\ x_{N1}\vec{e_1} + x_{N2}\vec{e_2} + \cdots + x_{Nr}\vec{e_r} \end{array} \tag{2.1}$$

The data can be presented in matrix format as:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1r} \\ x_{21} & x_{22} & \cdots & x_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nr} \end{pmatrix} \tag{2.2}$$

The $\vec{e_i}$ are the vectors generators of a subspace V. The aim being, to find a new basis vector $\vec{(y)}$ that defines a new subspace retaining the maximum amount of information from the original dataset:

$$\begin{array}{c} y_1 = w_{11}x_1 + w_{12}x_2 + \cdots + w_{1r}x_r \\ y_2 = w_{21}x_1 + w_{22}x_2 + \cdots + w_{2r}x_r \\ \vdots \\ y_m = w_{m1}x_1 + w_{m2}x_2 + \cdots + w_{mr}x_r \end{array} \tag{2.3}$$

Where $m \leq r$ and the x_i values the mean of each of the r variables across the N samples, that is $x_i = \frac{x_{1i}+x_{2i}+\cdots+x_{Ni}}{N}$. If we take $\mu_y = E(y)$ as the expected value of y then it follows that:

$$\mu_y = E(W^T X) = W^T E(X) \tag{2.4}$$

The covariance matrix of y is then:

$$C_y = E\{(y - \mu_y)(y - \mu_y)^T\} = W^T C_X W \tag{2.5}$$

To find the subspace with the maximum variability in the data, we calculate the covariance matrix of y (C_y), imposing a constraint of orthonormality:

$$W^T W = I \tag{2.6}$$

Which leads to:

$$(C_x - \lambda I)W = 0 \tag{2.7}$$

The problem has now been reduced to calculating the eigenvectors of C_X. The eigenvectors associated with the most significant values are then selected to form the subspace where the data have the greatest variability. To choose the number of principal components, we may use an index of variability defined as:

$$\frac{\sum_{i=1}^{l} \lambda_i}{\sum_{i=1}^{m}} \geq v \tag{2.8}$$

Where v is the amount of information to be retained. A value of 1 indicating that we retain all the eigenvalues.

2.3 Partial Least Squares (PLS)

Multi-variable linear regression (MLR), although it can handle a large number of descriptor variables, only works well for a large number of data points (more data-points than variables). When the number of data points is close to or smaller than the number of descriptor variables, MLR loses its prediction abilities as it will over-fit the data [14].

Principal Component Regression (PCR) captures the maximum variance in the data while MLR achieves maximum correlation between the explanatory variables and the target matrix. Partial Least Squares (PLS) tries to do both by maximising the covariance between the explanatory variables and the target matrix [15]. It requires the addition of weights to maintain orthogonal scores. The factors are calculated sequentially by projecting the target matrix Y through the explanatory matrix X. As noted in [15], there are a number of algorithms to compute the PLS of a dataset.

One possible algorithm for PLS is described in [17] and is presented again here. The input variables can again be represented in a matrix X as in Equation 2.2, the output matrix Y being similarly constructed

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ y_{21} & y_{22} & \cdots & y_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{Nq} \end{pmatrix} \tag{2.9}$$

The PLS technique then iteratively extracts factors from both X and Y such that the covariance between the extracted factors is maximised. This technique works with multivariate response variables ($q > 1$) but for our purpose we consider a vector Y ($q = 1$). The idea is to find a linear decomposition of X and Y such that $X = TP^T + E$ and $Y = UQ^T + F$, where T is an $(N * p)$ matrix of X-scores, U an $(N * p)$ matrix of Y-scores, P a $(r * p)$ matrix of X-loadings, Q a $(1 * p)$ vector of Y-loadings, E an $(N * p)$ matrix of X-residuals and F a $(N * 1)$ matrix of Y-residuals.

The decomposition is finalised to maximise the covariance between T and U. The number of extracted factors (p) depends on the rank of X and Y. One of the extraction methods is the eigenvalue decomposition algorithm. Each extracted x-score are linear combinations of X. For example, the first extracted factor a of X is of the form $a = Xw$ where w is the eigenvector of the first eigenvalue of X^TYY^TX. Similarly, the first y-score is $b = Yc$, where c is the eigenvector of the first eigenvalue of Y^TXX^TY. Note that X^TY denotes the covariance of X and Y. Now we deflate the X and Y matrices:

$$\begin{aligned} X_1 &= X - aa^TX \\ Y_1 &= Y - aa^TY \end{aligned} \tag{2.10}$$

The above process is repeated until all latent factors a and b have been extracted, that is until X is reduced to a null matrix.

As described in [16], the method is straightforward and reasonable in describing prediction equations. However, [18] highlights the main strengths and limitations of the method. The salient issues being of a higher risk of overlooking some correlations and a high sensitivity to the relative scaling of the descriptor variables.

References

[1] W. S. Noble and M. J. MacCoss, "Computational and statistical analysis of protein mass spectrometry data," *PLOS Computational Biology*, vol. 8, Jan. 2012.

[2] A. I. Nesvizhskii, "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomic," *Journal of Proteomics*, vol. 73, no. 11, pp. 2092–2123, 2010 [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1874391910002496

[3] European Bioinformatics Institute, "What is proteomics?" Aug-2011. [Online]. Available: https://www.ebi.ac.uk/training/online/course/proteomics-introduction-ebi-resources/what-proteomics. [Accessed: 09-Feb-2018]

[4] W. M. Niessen and D. Falck, "Introduction to mass spectrometry, a tutorial," in *Analyzing biomolecular interactions by mass spectrometry*, Wiley-VCH Verlag GmbH & Co. KGaA, 2015, pp. 1–54 [Online]. Available: http://dx.doi.org/10.1002/9783527673391.ch1

[5] "Mass spectrometry introduction." [Online]. Available: http://www.chem.pitt.edu/facilities/mass-spectrometry/mass-spectrometry-introduction. [Accessed: 19-Feb-2018]

[6] M. Spivak, "Analysis of mass spectrometry data for protein identification in complex biological mixtures," PhD thesis, Department of Computer Science, New York University, 2010.

[7] M. Hilario and K. A., "Approaches to dimensionality reduction in proteomic biomarker studies." *Briefings in Bioinformatics*, Mar. 2008 [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18310106/. [Accessed: 16-Feb-2018]

[8] D. L. Sampson, T. J. Parker, Z. Upton, and C. P. Hurst, "A comparison of methods for classifying clinical samples based on proteomics data: A case study for statistical and machine learning approaches," *PLoS ONE*, vol. 6, no. 9, 2011 [Online]. Available: https:

//www.ncbi.nlm.nih.gov/pmc/articles/PMC3182169/. [Accessed: 16-Feb-2018]

[9] J. M. Escaño, F. Dorado, and C. Bordons, "PCA based pressure control of a gas mixing chamber," in *2009 ieee conference on emerging technologies factory automation*, 2009, pp. 1–6.

[10] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37–52, 1987 [Online]. Available: http://www.sciencedirect.com/science/article/pii/0169743987800849

[11] M. Haenlein and A. M. Kaplan, "A beginner's guide to partial least squares analysis," *Understanding Statistics*, vol. 3, no. 4, pp. 283–297, 2004 [Online]. Available: https://doi.org/10.1207/s15328031us0304_4

[12] P. Geladi and B. R. Kowalski, "Partial least-squares regression: A tutorial," *Analytica Chimica Acta*, vol. 185, pp. 1–17, 1986 [Online]. Available: http://www.sciencedirect.com/science/article/pii/0003267086800289

[13] J. M. Escaño, "Fuzzy model predictive control. complexity reduction by functional principal component analysis," PhD thesis, Departamento de Ingenieria de Sistemas y Automática, Universidad de Sevilla, 2015 [Online]. Available: https://idus.us.es/xmlui/bitstream/handle/11441/32683/tesis_completa.pdf?sequence=1&isAllowed=y. [Accessed: 15-Feb-2018]

[14] R. D. Tobias, "An introduction to partial least squares regression," 1995 [Online]. Available: https://stats.idre.ucla.edu/wp-content/uploads/2016/02/pls.pdf. [Accessed: 12-Feb-2018]

[15] B. M. Wise, "Properties of partial least squares (pls) regression, and differences between algorithms." Eigenvector Research Incorporated [Online]. Available: http://www.eigenvector.com/Docs/Wise_pls_properties.pdf. [Accessed: 15-Feb-2018]

[16] P. H. Garthwaite, "An interpretation of partial least squares," *Journal of the American Statistical Association*, vol. 89, no. 425, pp. 122–127, 1994 [Online]. Available: http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1994.10476452

[17] S. Maitra and J. Yan, "Principle component analysis and partial least squares: Two dimension reduction techniques for regression," in *Discussion Paper Program*, 2008 [Online]. Available: https://www.casact.org/pubs/dpp/dpp08/08dpp76.pdf. [Accessed: 19-Feb-2018]

[18] R. D. Cramer, "Partial least squares (pls): Its strengths and limitations," *Perspectives in Drug Discovery and Design*, vol. 1, no. 2, pp. 269–278, Dec. 1993 [Online]. Available:

https://doi.org/10.1007/BF02174528