AMSI VACATION RESEARCH SCHOLARSHIPS 2019–20

EXPLORE THE MATHEMATICAL SCIENCES THIS SUMMER

Feature Extraction of Wheat Properties from Peptide Data

Timothy E Chapman

Supervised by Prof. Inge Koch and Dr. Lyron Winderbaum

University of Western Australia

Vacation Research Scholarships are funded jointly by the Department of Education and the Australian Mathematical Sciences Institute.

The ARC Centre for Plant Energy Biology (Director Harvey Millar) is concerned with the energy efficiency of plants in harsh environments. This project concentrates on one of their current projects involving peptide and protein measurements arising from genotypes in wheat. Data measured from wheat grown under two different environments have been analysed, with the aim of identifying factors related to the extensibility property. Environment has been found to have a significant impact on the physical characteristics. Canonical correlation analysis (CCA) has been applied to identify 231 oligo-peptides of interest from a collection of 13,566. The peptides identified as being the most important vary between environments, but a common overlap was found.

1 Introduction

Wheat is the world's most prevalent food source, accounting for 18% of the calories, and 20% of the protein consumed by humans (according to UN Food and Agricultural Organization (2013)), and provides a number of crucial vitamins and minerals.

Linking properties of wheat to its genome allows wheat breeders to develop crops with improved disease resistance, growth, and other attributes. There is already a large body of work evaluating crop resistance to pests and environmental conditions and for improving growth (Landjeva et al. 2009; Zegeye et al. 2014). However, another aspect worth improving is the quality of wheat flour. *Extensibility* —the stretching property of dough— is of key interest owing to its importance in bread and pasta (Rosada 2004). An improved understanding of what influences extensibility can deliver tremendous rewards in developing better crops.

Molecular information, frequently genetic and protein data, typically falls into the category of high-dimension low sample size data. This type of data necessitates a different approach, as many classical techniques are based on a small number of variables, or hold true as the number of samples becomes extremely large.

Multivariate analysis of the wheat genome, and observed traits in different environments, such as that done by Caffe-Treml et al. (2011), is required for the high-dimension low sample size data presented here. While a variety of techniques have been tried (Verbyla et al. 2012),

this analysis has used principal component analysis (PCA) and canonical correlation analysis (CCA). CCA in particular has been used successfully in other studies such as by Tang et al. (2012). As is pointed out by Naylor et al. (2010) this allows for the identification of complex relationships between genes that pairwise univariate techniques are unable to detect.

The genome provides information on how wheat varieties can behave. However, under different conditions wheat will behave very differently, as such with our current understanding, the genome alone provides incomplete information. Determining which genes actually get expressed is a difficult question the field of epigenetics is devoted to. An alternative approach is to examine the proteins transcribed from the genes. This project takes the second approach and studies the protein fragments identified in flour samples from different wheat varieties in two environments.

1.1 A brief note on the format used

The data analysis method warrants mentioning because it is sufficiently different from standard methods. A single document was used for all code, data analysis, and observations. This plain-text document uses the Org mode syntax, as laid out by Schulte et al. (2012) which (using a supporting code editor) enables an approach known as 'literate programming', and reproducible research. This document (notebook.org) is attached to this PDF, as are the data files (1 and 2).

1.2 Statement of Authorship

All of the techniques used herein are well established. PCA and CCA are graduate level topics which I learnt for this project, with some modifications to CCA to suit my data. Some of the notation and methodology is non-standard and based on sections one to three of Koch (2014). The content of notebook.org, results, and this report are all my own work.

I wish to acknowledge Inge Koch who was instrumental in the development of my understanding, ARC and Harvey Millar for providing the data, and Lyron Winderbaum who cleaned the data sets and provided assistance with the nuances of the R programming language.

2 Characteristic data

2.1 Pairwise correlation

Two data sets were used in this project. Both contained measurements made on 34 wheat varieties, grown in two environments (with one result missing), forming 67 samples.

The first data set, henceforth referred to as the 'characteristic' data, contains measurements of what may be considered to be the physical properties of the wheat and food products made from it — properties such as water content, starch damage, and extensibility. There are 87 such parameters.

For the planned analysis, we *centred* the data, such that the mean of each variable is zero. Since the variables have several different units, the large range in variances (as seen in figure 1) makes comparisons difficult. Thus it also made sense to scale the standard deviation of each variable to unity.

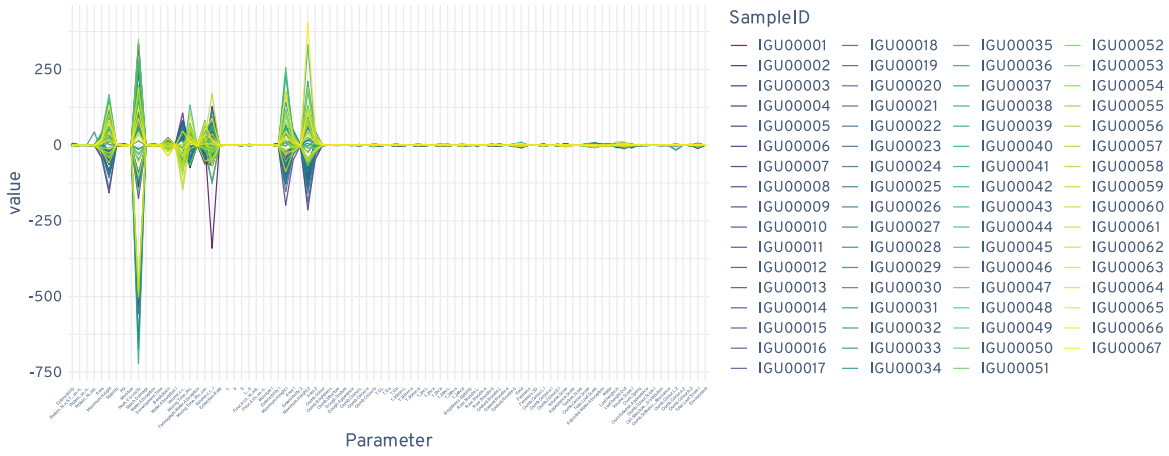

Figure 1: Parallel coordinate plot of characteristic data, centred and unscaled. Exhibiting a wide range of values, necessitating normalisations to unity.

2.2 Removal of spurious variables

To visualise the strength of the correlation between pairs of variables in the characteristic data set, a matrix was constructed in which the rows and columns correspond to variables, and each entry holds the squared correlation coefficient (R^2) of the respective pair of variables. R^2

is used rather than R as we are more interested in the strength than the sign. These correlation strengths are visualised in figure 2.

This figure reveals another issue with the characteristic data — a few pairs of variables for which one variable is effectively a linear transformation of the other, visible as the bright yellow cells. These redundant variables were removed by identifying those pairs with R^2 values above a high threshold (0.99), and removing the first variable in the pair from the data set, resulting in the removal of six variables (see appendix A.1). Two more variables have been removed as they have a constant value.

Inspecting the highly correlated clumps, one notices that the variables are generally different measurements of the same feature. Most of the pairs exhibit a very low level of correlation. Of interest are the exceptions that correlate to extensibility, which is the leftmost column.

2.3 Principal Component Analysis

2.3.1 Introduction to the technique

To explore the importance of the variables within the characteristic data set *Principal Component Analysis* (PCA) was applied, revealing the influence of the environment on the data.

Principal component analysis is a method which decomposes the covariance to summarise the data in a smaller number of variables —called principal components— without loosing any meaningful information. Each principal component (PC) is a linear combination of the original variables, and they are uncorrelated which makes them convenient to work with. The maximum number of components is the rank of the data.

Each PC has a *score* which is the projection of the centred data onto the principal component. The first principal component is such that it has maximal variance, and the second will have maximum variance with the restriction of orthogonality to all previous components, and so forth.

Principal components can be determined from the sample *covariance matrix* S. For a collection of random samples $\mathbb{X} = [\mathbf{X}_1\mathbf{X}_2 \dots \mathbf{X}_n]$ where each $\mathbf{X}_i$ is a random vector, the covariance matrix S is given by

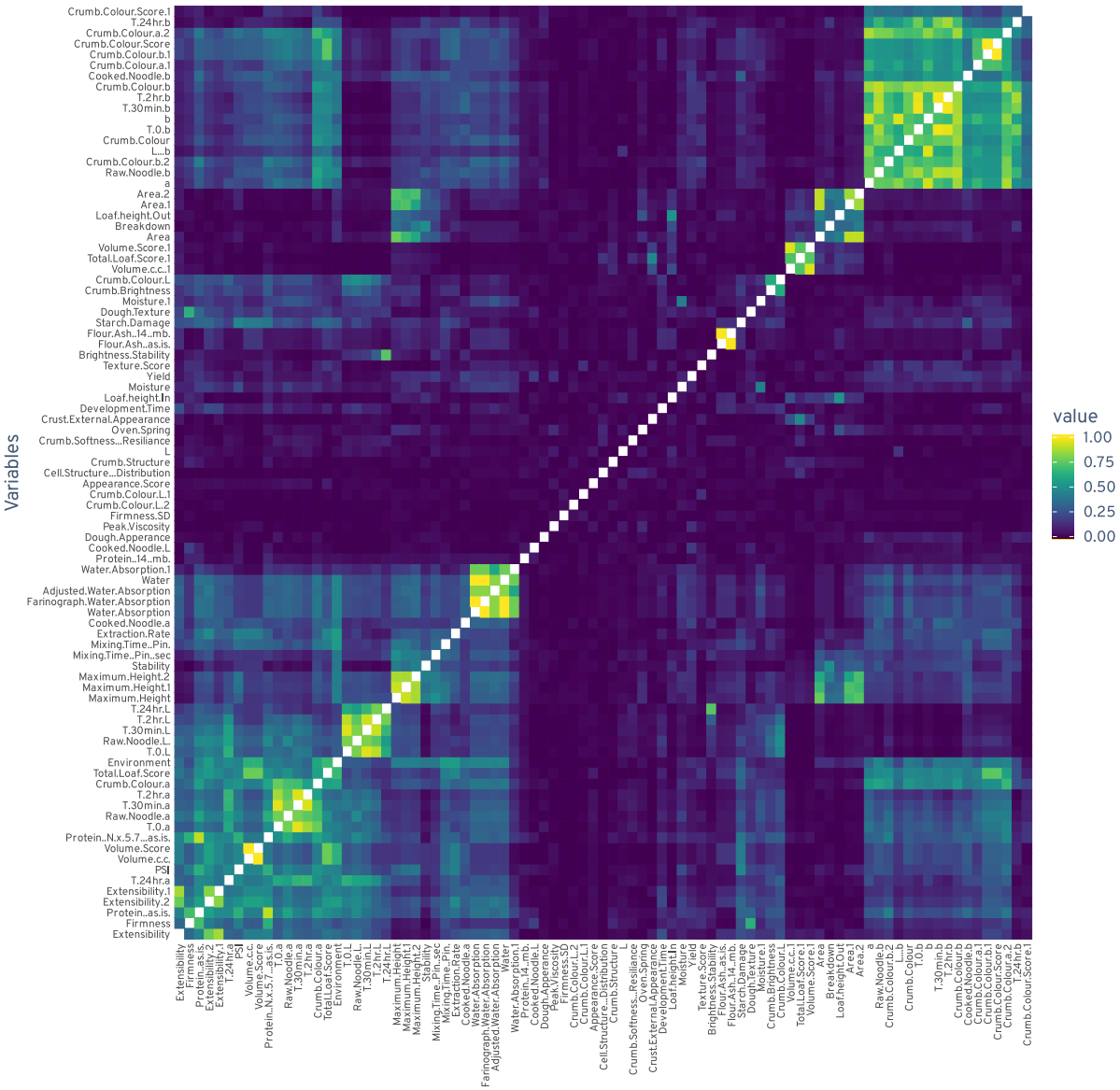

Figure 2: Pairwise correlation matrix for characteristic data, colour coded to indicate correlation strength (R^2 value). Correlations range from totally uncorrelated (dark purple) to perfectly correlated (yellow).

$$S = \begin{pmatrix} \mathrm{Cov}(\mathbf{X}_1, \mathbf{X}_1) & \mathrm{Cov}(\mathbf{X}_1, \mathbf{X}_2) & \dots & \mathrm{Cov}(\mathbf{X}_1, \mathbf{X}_d) \\ \mathrm{Cov}(\mathbf{X}_2, \mathbf{X}_1) & \mathrm{Cov}(\mathbf{X}_2, \mathbf{X}_2) & \dots & \mathrm{Cov}(\mathbf{X}_2, \mathbf{X}_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\mathbf{X}_d, \mathbf{X}_1) & \mathrm{Cov}(\mathbf{X}_d, \mathbf{X}_2) & \dots & \mathrm{Cov}(\mathbf{X}_d, \mathbf{X}_d) \end{pmatrix}$$

where Cov is the covariance function. By *centering* the data (shift such that $\overline{\mathbf{X}_i} = 0$), which we denote as $\mathbf{X}_{\text{cent}}$ the sample covariance matrix with n samples is conveniently $S = \frac{1}{n-1}\mathbf{X}_{\text{cent}}\mathbf{X}_{\text{cent}}^{\mathsf{T}}$.

To find the components the *spectral decomposition* of S can be used, which requires the matrix to be expressed in the form $S = \hat{\Gamma}\hat{\Lambda}\hat{\Gamma}^{\mathsf{T}}$. Here $\hat{\Gamma}$ contains the eigenvectors of S along its columns, and is an estimate of the population value Γ. $\hat{\Lambda}$ is a diagonal matrix which consists of the d non-zero eigenvalues of S arranged in decreasing order. By definition the k^{th} principal component is given by $\hat{\Gamma}_k^{\mathsf{T}}\mathbf{X}_{\text{cent}}$.

2.3.2 Application to Characteristic data

Applying PCA to the characteristic data results in 67 PCs (as 67 is the rank of the data).

The variance accounted for by each PC decreases rapidly as seen in figure 3. Hence for the purposes of this analysis, only the first few PCs warrant consideration.

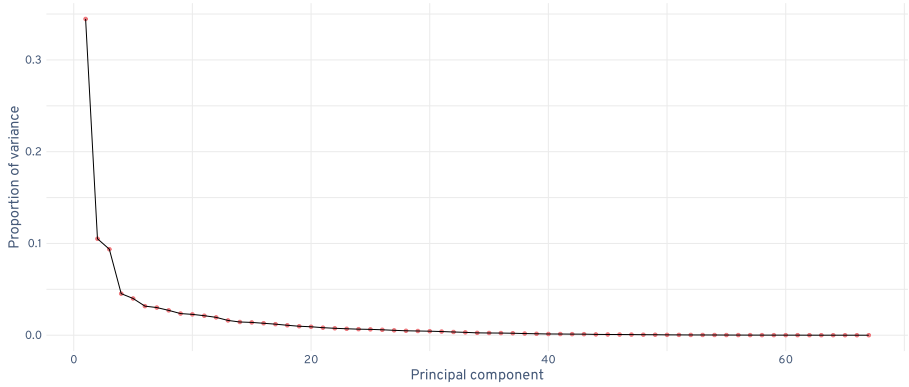

Figure 3: Proportion of variance accounted for by each principal component

Plotting the PCs against each other reveals the impact of the environment, seen in figure 4 where the samples are coloured based on their environment.

In figure 4, the environment clearly partitions the PC1 (first PC) scores, and PC1 has a much larger contribution to the total variance than the following PCs. This observation suggests

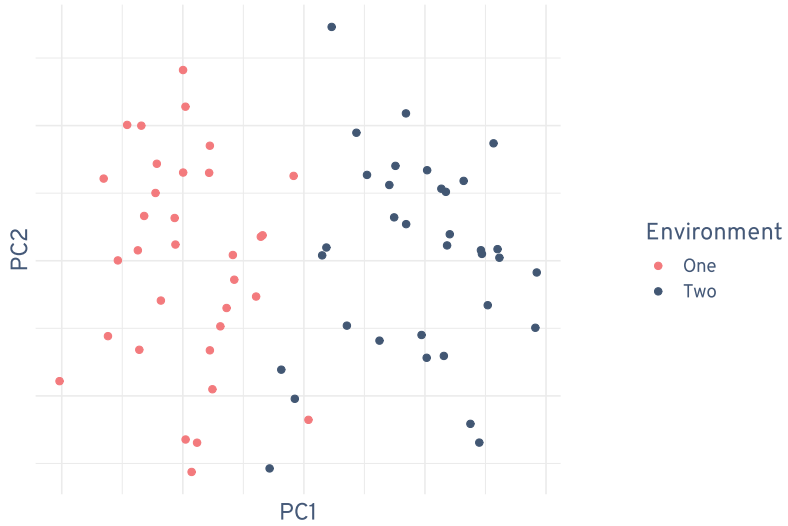

Figure 4: PC1 against PC2, where each sample produces a point by the linear combination of its variables with the weights given by the PCs; coloured by environment.

that environment has a significant impact on the rest of the variables, and as such it makes sense to consider the two environments separately.

2.4 Identification of characteristic variables related to extensibility

For the purpose of this investigation, not all the variables within the characteristic data set are of interest. To produce more relevant results the variables used have been restricted to those that seem related to extensibility. To determine which variables are most correlated to extensibility, linear modelling has been applied. Extensibility (y) is set as the 'response' variable, and represented as a linear combination of n 'predictor' variables (x_i), with undetermined weightings (α_i).

$$y = \sum_i \alpha_i x_i + \epsilon$$

We then seek to minimise $\sum_i \|y_i - \hat{y}_i\|^2$ (with y as the response variable, and $\hat{y}$ as the estimate) which resolves to n partial derivatives.

The R program has a built-in function `lm` for fitting linear models. This function used 67 variables to fit the 67 data points. This is a clear example of overfitting, it is as trivial as fitting a n^{th} degree hyperplane to $n+1$ data points, and hence is not a useful result.

In order to determine a more useful selection of variables, a technique known as *Forward Stepwise Selection* was applied. In this technique variables are added to the model one at a time. All variables are tested then the variable which contributes the most to the correlation coefficient are added (as described in Hastie et al. (2009) §3.3.2). This effectively gives the minimal set of variables needed to reach a given correlation coefficient.

As suggested by the partitioning of PC1 by environment, the results of this process were very different for the two environments, as seen in figure 5. The cumulative R^2 value rose more rapidly for environment one, and different variables were used for each environment.

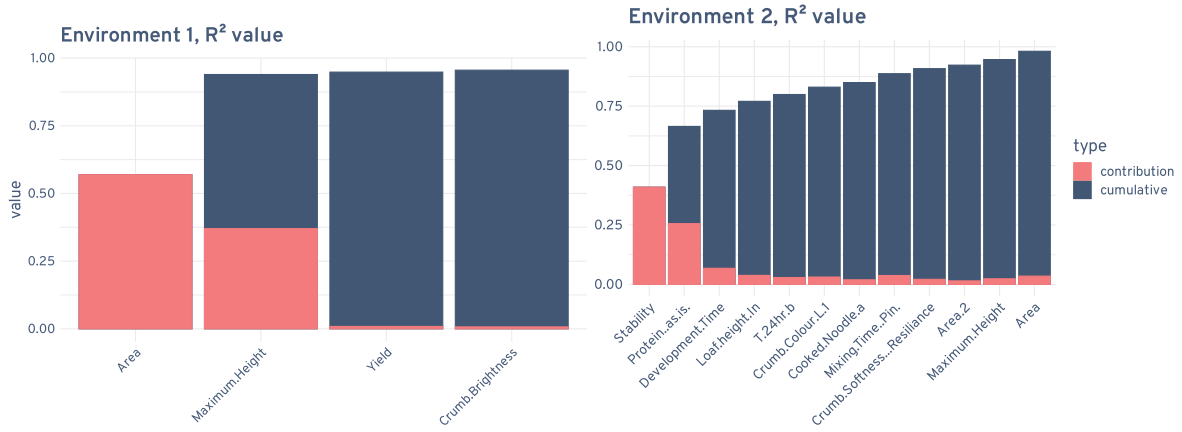

Figure 5: Results of Forward Stepwise Selection with characteristic data of both environments

Combining the variables identified in figure 5 a set of 15 variables which may be related to extensibility is obtained. This subset is a considerable reduction from the initial 79 variables.

3 Relation to peptide fragments

The second data can be considered to represent some of the molecular properties of the wheat; it is the assay of a set of unique protein fragments by mass spectrometry. The data provided contains 13,566 such fragments.

3.1 Can PCA produce a similar partition by environment?

Considering the result of PCA in section 2.3.2 PCA is also applied to the peptide data. Here PC1 accounts for a particularly large proportion of the variance — over half, and the decay in PC contribution to variance is more striking, as seen in figure 6. PC1 clearly dominates, however,

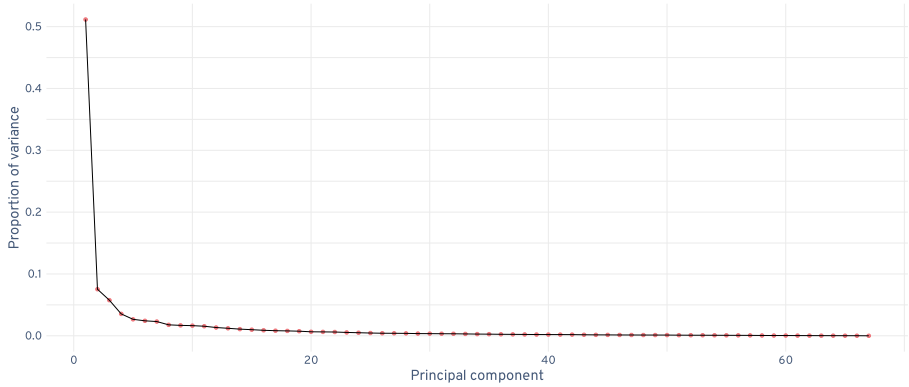

Figure 6: Proportion of variance individual PCs account for

it doesn't exhibit the partitioning due to environment observed with the characteristic data in figure 4. Instead, in figure 7 the clusters for each environment overlap in the top right corner in a linear fashion. The outlier in the bottom left of figure 7 (IGU00002) skews the cluster away

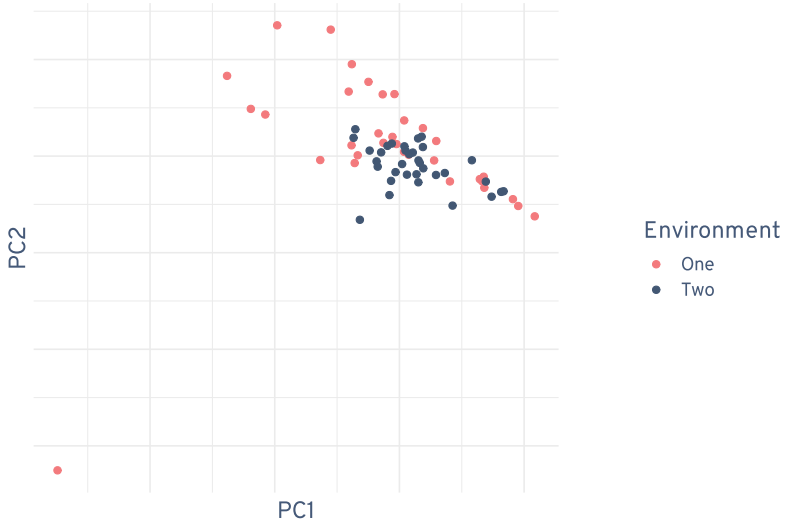

Figure 7: PC1 against PC2 for the peptide data, coloured by environment

from the centre. While the cluster for the first environment appears to have a greater spread than the second, there is no partitioning similar to that seen with the characteristic data in figure 4.

3.2 Canonical Correlation Analysis

3.2.1 The canonical correlation matrix

The two data sets contain two sets of observations made on the same samples. To find correlations between the two data sets this method considers linear combinations of the variables from each data set. Canonical correlation analysis (CCA) provides a method which uses common samples to identify relationships between the sets of variables. If one data set contains a single variable then CCA provides the same result as multivariate linear regression. The much more interesting case is when both data sets contain multiple variables, where CCA provides optimal combinations of variables for the data sets and considers every variable as both a response and predictor. In this manner, CCA is much more holistic than linear regression.

The well-known Pearson correlation coefficient is given by

$$\operatorname{Corr}(x_1, x_2) = \frac{\operatorname{Cov}(x_1, x_2)}{\sqrt{\operatorname{Var}(x_1)\operatorname{Var}(x_2)}} = \operatorname{Cov}(x_1, x_1)^{-1/2}\operatorname{Cov}(x_1, x_2)\operatorname{Cov}(x_2, x_2)^{-1/2}$$

With two data sets, denoted $\mathbb{X}^{[1]}$ and $\mathbb{X}^{[2]}$, in a manner very similar to finding the Pearson correlation coefficient, a matrix can be found which connects our two data sets — the sample *Canonical Correlation Matrix*. As in definition 3.4 of Koch (2014)

$$\hat{C} = S_{11}^{-1/2} S_{12} S_{22}^{-1/2}$$

where S_{ij} is (as given earlier in section 2.3.1) the sample covariance matrix, just between two data sets, and hence is given by $\frac{1}{n-1}\mathbb{X}_{\text{cent}}^{[1]}\mathbb{X}_{\text{cent}}^{[2]\,\mathsf{T}}$.

With this matrix, $\hat{C}$ the aforementioned optimal combinations of variables can be determined using a technique known as Singular Value Decomposition (SVD).

3.2.2 Singular value decomposition

An invertible, square matrix A can be diagonalised into the form $A = EDE^{-1}$, where E contains A's eigenvectors as columns, and D is a diagonal matrix with A's eigenvalues along the diagonal. SVD can be thought of as a way to generalise this decomposition for any matrix. The $m \times n$ matrix $\hat{C}$ with rank r can be decomposed into the form $\hat{P}\hat{\Upsilon}\hat{Q}^{\mathsf{T}}$, seen in figure 8.

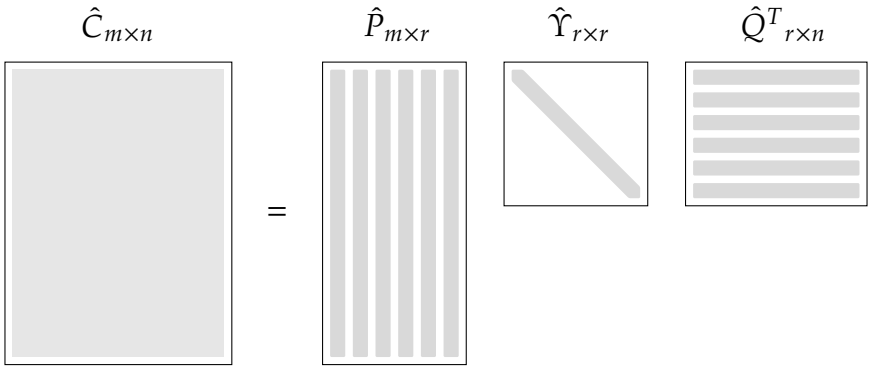

Figure 8: A diagramatic representation of singular value decomposition

The columns of $\hat{P}$ and $\hat{Q}$ are orthogonal and act as eigenvectors. Hence the columns of $\hat{P}$ are referred to as the *left eigenvectors*, and those of $\hat{Q}$ as the *right eigenvectors*. Similarly, $\hat{\Upsilon}$ is a diagonal matrix the values of which are arranged in decreasing order and referred to as the *singular values*. This form is useful since it possesses many of the properties of diagonalisation. It is particularly useful in the case of CCA, specifically the SVD of $\hat{C}$,

What is useful in this case is that $\hat{P}$ and $\hat{Q}$ can be used to find the optimal weights of the variables, and the values along $\hat{\Upsilon}$ give the strength of the correlation. In particular making use of the columns of $\hat{P} = [\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_r]$ and $\hat{Q} = [\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_r]$. Each combination gives a pair of *canononical correlation scores* $\mathbf{U}_k = \hat{p}_k^\top \mathbf{X}_S^{[1]}$ and $\mathbf{V}_k = \hat{q}_k^\top \mathbf{X}_S^{[2]}$, where $\mathbf{X}_S$ is the *sphered* matrix $\mathbf{X}_S = S^{-1/2}\mathbf{X}_{\text{cent}}$.

The k^{th} best combination of variables is given by $\hat{\phi}_k = S_{11}^{-1/2}\hat{p}_k$ and $\hat{\psi}_k = S_{22}^{-1/2}\hat{p}_k$, where $\hat{\phi}_k$ and $\hat{\psi}_k$ give the weightings of the variables of $\mathbf{X}^{[1]}$ and $\mathbf{X}^{[2]}$ respectively. The plot of the combinations for the first data set against the second is known as the k^{th} *scores plot*, with each sample producing a point.

3.3 Identifying peptide fragments of interest

As laid out in section 3.2, CCA was performed on the peptide data set ($\mathbf{X}^{[1]}$) against a subset of the characteristic data set ($\mathbf{X}^{[2]}$), specifically extensibility and the 15 correlated variables identified in section 2.4. This resulted in 16 score plots, the first and last of which are seen in figure 9.

Every score plot exhibits perfect correlation, thus since extensibility is of foremost importance,

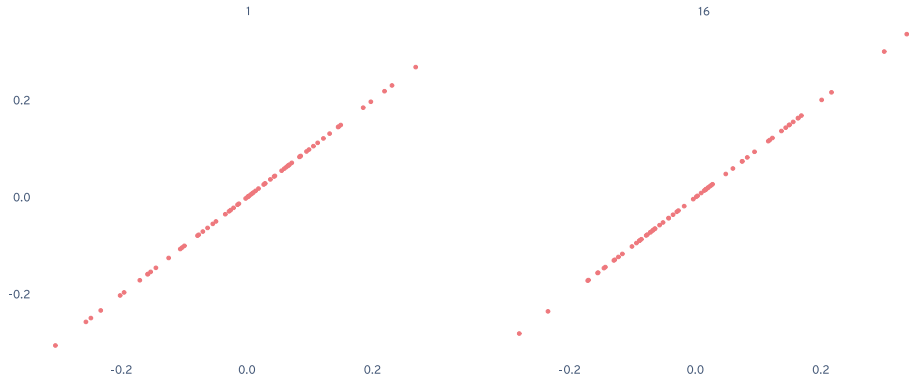

Figure 9: First and last score plot from CCA of the peptide fragment data against a subset of 16 variables from the characteristic data

the set of combinations in which extensibility has the largest impact is selected. To identify this combination a *relative weighting* was devised. Considering the set of weighs as a vector each individual weight is divided by the magnitude of that vector, and scaled it according to the expected magnitude should the weights be equally distributed across all the variables in the original data. The resulting transformation for a set of weights ϕ_i is given by,

$$\phi_{ij} \mapsto \sqrt{n}\frac{\phi_{ij}}{\|\phi_i\|}$$

The relative weighting of extensibility in each set of combinations is shown in figure 10.

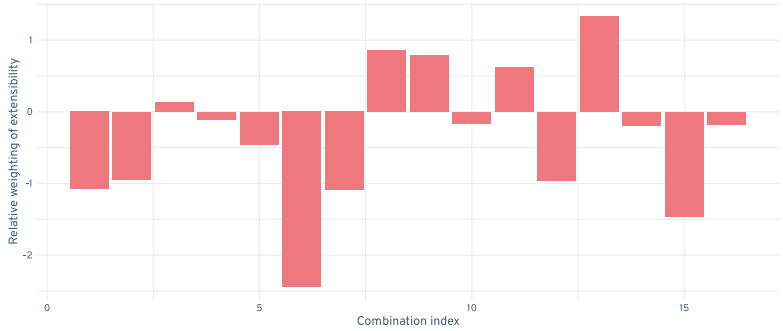

Figure 10: The relative weighting of extensibility across each combination from CCA of the peptide fragment data against a subset of 16 variables from the characteristic data

The sixth combination has the largest weighting for extensibility. To identify the most interesting peptide fragments, their relative weights in the selected combination were considered. The distribution of these relative weights seen in figure 11 was evaluated, and a cut-off value of five was chosen as it produced a reasonable number of peptides over the threshold.

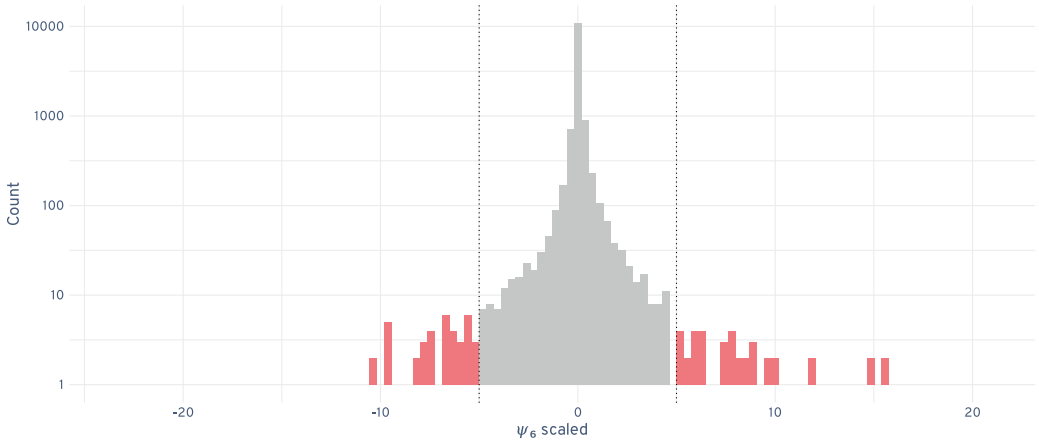

Figure 11: Histogram of relative weightings of peptide fragment variables from the sixth combination of variables. Shown in red are the variables over the cut-off (5).

From this, the 231 peptides which have the largest impact on the combination of variables in which extensibility had the greatest impact have been determined. This subset contains a mere 1.7 % of the peptide fragment variables.

To investigate this subset a pairwise correlation plot was produced from those 231 variables, and another from a random selection of the same number of variables. The result can be seen in figure 12. The plot for the variables of interest exhibits more, larger clusters with high correlation coefficients — which would be expected if groups of peptide fragments in the subset of interest are derived from the same protein, or proteins expressed together. This corresponds with what would be expected if the subset is relevant, which although not quantified, can be considered to be a good sign for the subset.

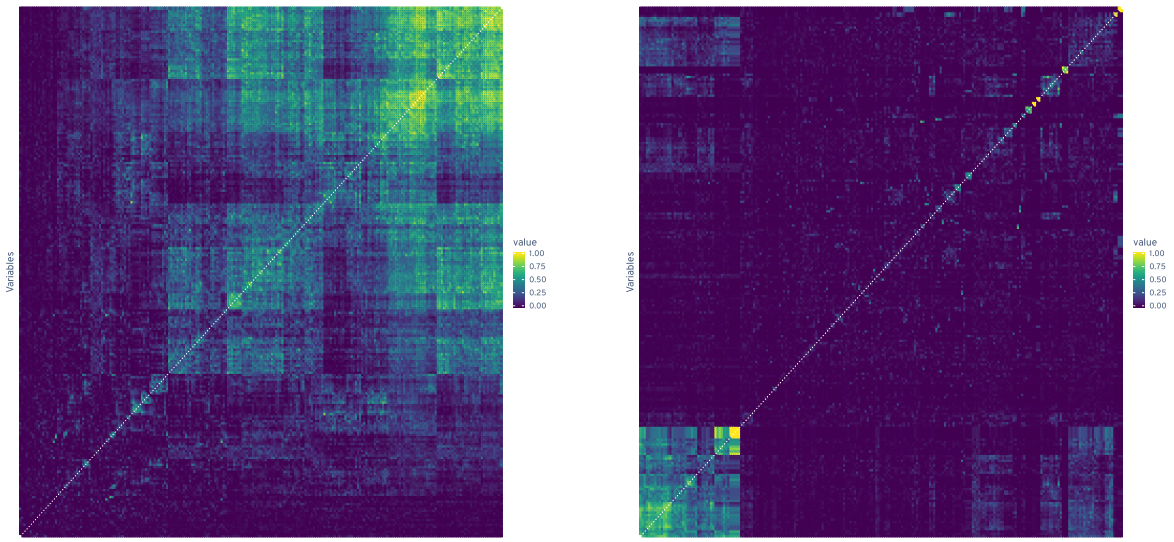

Figure 12: Pairwise correlation plot for 231 interesting peptide variables (**left**), and a random selection of 231 peptide variables (**right**). Reordered to maximise clustering.

3.4 Environmental effect on which peptide fragments are the most interesting

As was demonstrated in section 2.3.2 the effect of environment explains a substantial proportion of variance in the characteristic data. To examine how this affected which peptide fragments were identified as being most significant, the procedure of section 3.3 was repeated separately for each environment. The Jaccard index, which is a measure of similarity of two sets given by $\frac{|A \cap B|}{|A \cup B|}$, is calculated with A and B as the two environments, for a growing selection of the top k peptide variables. This value is plotted with the expected index (see appendix A.2) of a random selection of k peptides.

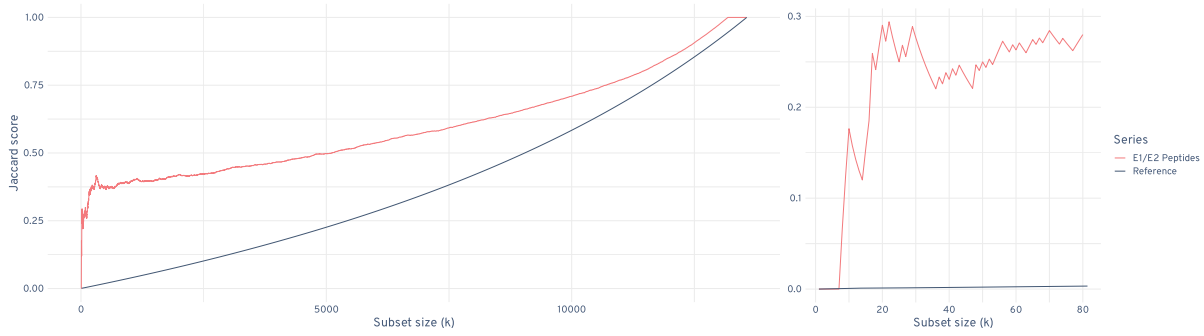

Figure 13: Development of the Jaccard index between the top k peptides fragments of the two environments, and the expected value for a random selection of k variables

Interestingly, figure 13 reveals the Jaccard index grows rapidly to around 0.3 by the first twenty variables, and then slowly climbs from there. Around a third to half of the peptides are shared between the top k lists for both environments.

3.5 Conclusions and further work

This investigation has identified the characteristic variables strongly related to extensibility, and through CCA has identified a subset of 231 peptide fragments (1.7 % of the original data set) which appear to affect the extensibility of wheat dough. The pairwise correlation plot of this subset exhibited groups that correlated well together, which is considered to be a good indication that a relevant selection has been identified.

Given the apparent importance of environment, it would be interesting to expand the data set to include more than two environments so that the effect could be further analysed.

It would be useful to gain a deeper insight into the underlying processes and the biological significance of the interesting peptide fragments. These results have already formed the basis for continuing investigation by the researchers who provided the data.

A Appendix

A.1 Characteristic data — removed variables

Variable to remove	Other variable	R^2
Water.Absorption	Water	0.999997
Flour.Ash..14..mb.	Flour.Ash..as.is.	0.992930
Volume.c.c.	Volume.Score	0.999999
Crumb.Colour.b.1	Crumb.Colour.Score	0.999999
Farinograph.Water.Absorption	Water	0.999997

A.2 The Jaccard similarity index

The Jaccard similarity index is defined as

$$\mathrm{Sim}_J(S,T) = \frac{|S \cap T|}{|S \cup T|}$$

In this report S, T are a random subset of $\{1, 2, 3, \ldots, n\}$. When the first k items out of n are picked from the two sets, the probability that all match is $\binom{n}{k}^{-1}$. For all but i to match, (a proportion of $\frac{k-i}{k+i}$) we would need i values from the $(k, n]$ range, which can happen $\binom{n-k}{i}$ ways, and for the i values to be arranged any way in the list, which can happen $\binom{k}{k-i}$ ways. It follows that the expected value of the Jaccard similarity index is given by,

$$\mathbb{E}[\mathrm{Sim}_J(S,T)] = \frac{1}{\binom{n}{k}} \sum_{i=0}^{k-1} \frac{k-i}{k+i} \binom{n}{k-i} \binom{n-k}{i}$$

References

Caffe-Treml, M., Glover, K. D., Krishman, G. G., Hareland, G. A., Bondalapati, K. D., & Stein, J. (2011). Effect of Wheat Genotype and Environment on Relationships Between Dough Extensibility and Breadmaking Quality. *Cereal Chemistry; St. Paul*, *88*(2), 201–208. Retrieved 2019, from http://search.proquest.com/docview/863261361/citation/E1452FA954A74CC4PQ/1

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. doi:10.1007/978-0-387-84858-7

Koch, I. (2014). *Analysis of multivariate and high-dimensional data*. OCLC: 872140281. New York: Cambridge University Press. Retrieved 2020, from https://doi.org/10.1017/CBO9781139025805

Landjeva, S., Lohwasser, U., & Börner, A. (2009). Genetic mapping within the wheat D genome reveals QTL for germination, seed vigour and longevity, and early seedling growth. *Euphytica*, *171*(1), 129. doi:10.1007/s10681-009-0016-3

Naylor, M. G., Lin, X., Weiss, S. T., Raby, B. A., & Lange, C. (2010). Using Canonical Correlation Analysis to Discover Genetic Regulatory Variants. *PLoS One*, *5*(5). doi:10.1371/journal.pone.0010395. pmid: 20485529

Rosada, D. (2004). Dough strength: Evaluation & techniques. *San Francisco Baking Institute: What's Rising*, 12. Retrieved from http://www.sfbi.com/pdfs/NewsF04a.pdf

Schulte, E., Davison, D., Dye, T., & Dominik, C. (2012). A Multi-Language Computing Environment for Literate Programming and Reproducible Research. *J. Stat. Soft.* *46*(3). doi:10.18637/jss.v046.i03

Tang, C. S., & Ferreira, M. A. R. (2012). A gene-based test of association using canonical correlation analysis. *Bioinformatics*, *28*(6), 845–850. doi:10.1093/bioinformatics/bts051

UN Food and Agricultural Organization. (2013). FAOSTAT. Retrieved 2019, from http://www.fao.org/faostat/en/#data/FBS

Verbyla, A. P., & Cullis, B. R. (2012). Multivariate whole genome average interval mapping: QTL analysis for multiple traits and/or environments. *Theor Appl Genet*, *125*(5), 933–953. doi:10.1007/s00122-012-1884-9

Zegeye, H., Rasheed, A., Makdis, F., Badebo, A., & Ogbonnaya, F. C. (2014). Genome-Wide Association Mapping for Seedling and Adult Plant Resistance to Stripe Rust in Synthetic Hexaploid Wheat. *PLOS ONE*, *9*(8), e105593. doi:10.1371/journal.pone.0105593