Original article for Food Chemistry

Application of sequential and orthogonalised-partial least squares (SO-PLS) regression to predict sensory properties of Cabernet Sauvignon wines from grape chemical composition

Jun Niimi[†‡*], Oliver Tomic[¶], Tormod Næs[§], David W. Jeffery[†], Susan E. P. Bastian[†], Paul K. Boss[‡]

[†]*School of Agriculture, Food and Wine, The University of Adelaide, PMB 1, Glen Osmond, SA 5064, Australia*

[‡]*CSIRO - Agriculture and Food, PMB 2, Glen Osmond, SA 5064, Australia*

[¶]*Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, 1432 Norway*

[§]*Nofima – Norwegian Institute of Food, Fishery and Aquaculture, Ås, 1431, Norway*

*****Corresponding author*

CONTACT DETAILS: jun.niimi@adelaide.edu.au, +61 8 8313 0284

Abstract

The current study determined the applicability of sequential and orthogonalised-partial least squares (SO-PLS) regression to relate Cabernet Sauvignon grape chemical composition to the sensory perception of the corresponding wines. Grape samples (n = 25) were harvested at a similar maturity and vinified identically in 2013. Twelve measures using various (bio)chemical methods were made on grapes. Wines were evaluated using descriptive analysis with a trained panel (n = 10) for sensory profiling. Data was analysed globally using SO-PLS for the entire sensory profiles (SO-PLS2), as well as for single sensory attributes (SO-PLS1). SO-PLS1 models were superior in validated explained variances than SO-PLS2. SO-PLS provided a structured approach in the selection of predictor chemical data sets that best contributed to the correlation of important sensory attributes. This new approach

presents great potential for application in other explorative metabolomics studies of food and beverages to address factors such as quality and regional influences.

Keywords: Multi-block data analysis; data orthogonalisation; grape; wine; sensory

1.0 Introduction

The field of metabolomics analysis is rapidly expanding in the quest to improve the holistic understanding of food and beverage composition in relation to nutrition, quality, safety, and authenticity (Wishart, 2008). Given that consumers are an important beneficiary of any improved knowledge of processes and practices, researchers continue to search for the compositional factors that contribute to flavour perception of products such as wine, which in turn influence consumer behaviour. Detailed studies of this nature generate substantial volumes of multiple data sets, which require suitable methods for data analysis to draw conclusions about complex natural phenomenon.

Fortunately, the field of chemometrics provides a range of multivariate statistical methods available for explorative analysis, interpretation and prediction. Chemometric methods also need to keep pace with the emerging trend involving the collection of multiple data sets obtained from advanced instrumental technologies with enhanced measurement resolution (and are therefore very information-rich). However, the large amounts of data that are generated pose a major challenge in the subsequent analysis to be able to interpret their meaning (Johnson, Ivanisevic, Benton, & Siuzdak, 2015).

Recent developments in partial least squares (PLS)-based analyses have involved extensions to multiple input data by using the PLS algorithm, including sequential and orthogonalised-PLS (SO-PLS) and parallel orthogonalised-PLS (PO-PLS) (Næs, Tomic, Mevik, & Martens, 2011). These techniques orthogonalise multiple data blocks, which maintain the integrity of each block and can account for their respective variation to the overall model. This may represent an advantage over joining multiple data sets together into one large table (concatenation) for analysis with PLS (multi-block-PLS) (Westerhuis, Kourti, & MacGregor, 1998). Specifically for the SO-PLS, data blocks of independent variables (X_1, X_2, etc.) are sequentially added to the analysis one at a time in succession to determine the progressive change in explained variance of the global response data (Y) (Næs, Tomic, Mevik, & Martens, 2011). Such new data analysis methods give a structured approach to the

analysis of highly complex data, thereby providing the best chance of properly modeling the phenomenon being studied.

Despite being prime tools for investigating the results of something as complex as human perception of foodstuffs, multi-block data analyses have been applied in only a few studies to describe sensory perception and consumer preferences. Perhaps this is mainly due to absence of such methods in standard commercial data analysis software that provide their access through graphical user interfaces. Multi-block-PLS analysis was applied to the study of aroma perception and release in cheeses in an attempt to determine the reasons behind large inter-individual differences in aroma release (Feron, Ayed, Qannari, Courcoux, Laboure, & Guichard, 2014). In this case, multiple predictor data sets were pre-processed prior to concatenation, the PLS algorithm applied, followed by multi-block redundancy analysis. Without taking the multi-block-PLS approach, mastication parameters as well as bolus characteristics of cheese would not have been identified as influencing aroma compound release in the mouth. PO-PLS was compared against conventional PLS for consumer preference mapping of flavoured water and jams (Måge, Menichelli, & Næs, 2012). Although the two analyses did not change in the output of the model such as the overall explained variance, the PO-PLS approach added further information in the contribution of predictor data blocks to the overall model and the number of components required for each block. Lastly, SO-PLS was applied to path modeling to determine how consumer demographics, purchase behaviour, and neophobia influenced each other (Menichelli, Almoy, Tomic, Olsen, & Naes, 2014). Many other applications of SO-PLS or PO-PLS to food and beverage research can be envisaged when diverse data sets need to be modelled (e.g., chemical, human sensory, biochemical), such as the case of working with grape and wine data.

Given the size and economic importance of the global wine industry, understanding the drivers of grape quality and how this translates into a finished wine that consumers appreciate is a fundamental requirement for winemakers. Although much research has elucidated grape chemical compounds that dictate some unique flavour characters in wines, there is still much to be understood concerning why wine taste the way they do. For instance, it is unknown why different wines from the same cultivar (i.e., cultivated variety, such as Cabernet Sauvignon) possess different sensory characteristics, as a result of the complicated interdependencies that occur between the chemical constituents within the grape berry, overlaid with the effects of microbial metabolism during winemaking. Chemometric treatment of data, in particular using multi-block data analysis methods, has the potential to

provide improved understanding of the grape chemical measures that best contribute to the variation in wine style as determined from instrumental and/or sensory profiles of the wines.

The objective of this study was to explore the applicability of SO-PLS to model the sensory characteristics of Cabernet Sauvignon wines as determined by human assessors using a suite of chemical measurements made on the grapes, and hence to determine the data blocks that most contribute to the models. In particular, focus will be on strategies for incorporating blocks of data when the number by far exceeds the standard size of 2-3 input blocks modelled by the approach. Focus will be on interpretation as well as prediction ability and how to assess reliability of the interpretations.

2.0 Materials and methods

2.1 Sampling

To encompass a range of compositional differences, Cabernet Sauvignon grapes were sampled from different vineyards within the following eight viticultural regions across South Australia during the 2013 vintage (number of vineyards sampled given in parentheses): Barossa Valley (2) (BV), Clare Valley (2) (CV), Coonawarra (4) (CWA), Eden Valley (2) (EV), Langhorne Creek (2) (LC), McLaren Vale (2) (McL), Riverland (9) (RVL), and Wrattonbully (2) (WBY). This provided a sample set of 25, with the Riverland having more vineyards sampled due to its comparatively larger sampling area.

In each vineyard site, grape bunches were randomly sampled throughout the vineyard block and were picked randomly from all areas of the canopy (Calderon-Orellana, Mercenaro, Shackel, Willits, & Matthews, 2014) to give a total parcel size of 60 kg. Two subsamples of grape berries (500 g each) were randomly taken from each sample parcel; one subsample was used for wet chemistry (see supplementary experimental section) and the other was snap frozen with liquid nitrogen and stored at -80 °C for compositional analysis. The pulp and skin of frozen grape samples were later separated from the seeds, and the frozen pulp and skin were homogenized, all the while under liquid nitrogen, in preparation for the various analyses. The remaining fresh grape parcels were processed into single wines, corresponding to each grape sample, and vinified identically using a small scale procedure (Niimi, Boss, Jeffery, & Bastian, 2017).

2.2 Chemical and sensory analyses

Twelve physical and chemical analyses were made on the grape samples (Table 1). For brevity, details of each analytical method are described in the Supplemental Information sections. Many of the methods have been described previously in literature and the reader is referred to those cited for the following measures of grapes; harvest measures, amino acids, non-targeted volatile compounds, total phenolics and total tannins, anthocyanins, detailed tannins, flavonols, and lipoxygenase enzyme activity pathway. Details for the remaining blocks of data including certain aspects of harvest measures, targeted and bound volatile compounds, colour, and fatty acid analyses are described in S-1 to S-5. The sensory analysis procedure was based on previous literature, however specific details pertaining to the current study are described in S-6. Sensory evaluations were conducted under the approval of The University of Adelaide's Human Research Ethics Committee (H-2014-057).

2.3 Data Analysis

2.3.1 Data pre-processing

The following pre-processing steps were used in the current study: (I) computing descriptive statistics for the variables to inspect the distribution of the data and identify possible severe outlier data points; and (II) analysing the data of each variable (both X- and Y-blocks) using one-way analysis of variance (ANOVA) to identify variables that significantly discriminated the wine samples ($p < 0.05$). Non-significant variables were removed from further analysis in an attempt to remove noise and ease the burden of computing the SO-PLS models (described further below); (III) analysis of each block with principal component analysis (PCA) to get an overview of the systematic variance in each block and how the variables contributed to the variance as well as visualize any outliers; (IV) computation of PLS2 models with Y- and one X-block at a time (totalling 12 PLS2 models) for an understanding of the predictive power of each X-block; (V) X-blocks that yielded PLS2 models with less than 10% validated explained variance were left out to remove noise from data. Further, single Y-variables with less than 10% validated explained variance when modelled using PLS1 with any of the X-blocks were also left out in another attempt to remove noise from the data.

Note that initially, all blocks except X_{01} consisted of 75 rows (for complete information on the block dimensions see Table 1) consisting of 3 replicates for each of the 25

measured wines. Prior to analysis as in (III), (IV), (V), and with SO-PLS, averages were computed across the 3 replicates, such that the number of rows of each block was decreased to 25. The order of the 25 wine samples was identical for all blocks. Moreover, all blocks were mean centred prior to analysis in (III), (IV), (V) and SO-PLS.

2.3.2 Analysis of multiple blocks using SO-PLS

The SO-PLS approach used has been generically described previously; for further information the reader is encouraged to read the following reference (Næs, Tomic, Mevik, & Martens, 2011). However, as the current paper will present two variants of SO-PLS, the distinction between single Y-variable SO-PLS and global SO-PLS (which analyses the entire Y-variables table) will be referred hereinafter as SO-PLS1 and SO-PLS2, respectively, following the conventional naming of the two PLS methods (i.e., PLS1 and PLS2).

A progressive model optimization approach was chosen, where X-blocks were added sequentially to the SO-PLS model (Menichelli, Almoy, Tomic, Olsen, & Naes, 2014) until there was no further reduction in root mean square error of cross validation (RMSECV) for Y-block. This is less likely to lead to over-fitting of the data, than finding the most optimal combination of X-blocks by model parameters using the global modeling approach (Menichelli, Almoy, Tomic, Olsen, & Naes, 2014). The reason for this is that the more possibilities there are, the higher the likelihood of obtaining a good result by chance. The progressive model optimization approach is described in detail below.

As a first step, the Y-block was fitted to each of the available X-blocks by computing one PLS2 model for each X-block. The models then were compared by means of lowest RMSECV for block Y and the appropriate number of components. The X-block with the most optimal parameters was selected to be the first X-block for subsequent SO-PLS2 models that will be computed in the following steps. The appropriate number of components found for the most optimal X-block at this point was set and held constant for the SO-PLS2 models.

As a second step, SO-PLS2 models with two X-blocks were calibrated. In each of these SO-PLS2 models the first X block and its number of components were fixed (as found in step 1 above) and the second X-block was one of the remaining X-blocks after step 1. This step determined the second X-block that most decreased RMSECV in Y-block, where the 2 block SO-PLS2 model with the lowest RMSECV was considered the most optimal model. The optimal number of components associated with the best second X-block as well as the

improvement in validated explained variance were also determined. Thus, the combination of decreased RMSECV, number of components, and the improved validated explained variance from the initial PLS2 were used as indicators that the 2 block SO-PLS2 model was better at describing the variation in block Y.

Upon determining an improved model with 2 block SO-PLS2 model, the number of components for the second block was set and held constant, as with the first X block. The process was repeated by adding a third X-block from the remaining X-blocks, making a 3 block SO-PLS2 model (still the number of components is fixed in the first two). For the data set below, the third block did not improve predictions and therefore only two blocks were considered throughout the study. It should be noted that although the present approach is more conservative than other selection strategies, prediction ability measures should be validated further with data from harvests of following years.

Once the final model was identified, the explained variance of Y was computed after the sequential addition of each X-block to show their progressive contribution to the total explained variance in Y. Moreover, cross validation-ANOVA (CV-ANOVA) (Indahl & Naes, 1998) was used to test whether incorporating the X-blocks is statistically significant. This method is based on comparing squared cross-validated residuals for different models using paired t-tests. More specifically, the cross-validated residuals are calculated for zero, one and two blocks in the model. One-block residual is compared with zero-block (using only the mean of Y as an estimate) residual and two-block residual with one-block residual. When more than two output variables are used, the sum of the cross-validated residuals is used in the t-test.

As the last step of the SO-PLS2 procedure, Y was predicted from the optimal model and subsequently analysed using principal components of prediction (PCP) (Langsrud & Næs, 2003). PCP implies that a PCA is first run for the predicted Y-values before these predicted Y-values are related to all the X-variables using regression analysis. The X-blocks are then standardized, put together in one block and then regressed onto the principal components of predicted Y. Scores and loading plots for both Y and combinations of X-blocks were then plotted for interpretation of the results.

The above progressive SO-PLS2 modeling procedure was further performed with SO-PLS1, where one Y-variable was modelled at a time with the purpose to investigate which variables in the X-blocks were responsible for high or low intensities of that specific

Y-variable, in this case a particular sensory attribute. As above with SO-PLS2 this approach was used to reduce the heavy computational burden as much as possible, as well as reducing chances of serious overfitting, as described above.

All analyses were performed using the Python programming language (Python version 3.5) utilizing the Python packages *numpy* (Peréz & Granger, 2007), *IPython*, (Oliphant, 2007), *pandas* (McKinney, 2010), and *statsmodels* (Seabold & Perktold, 2010). The Python implementation of SO-PLS was coded in-house.

3.0 RESULTS AND DISCUSSION.

3.1 Data pre-processing checks

Data were initially checked with descriptive statistics to determine the distribution of data in all blocks and from it, unusual distributions were not detected. Each data block was analysed using one-way ANOVA. Data blocks X_{02}, X_{04}, X_{05}, X_{09}, X_{11}, and Y contained non-significant variables and were further reduced to 24, 26, 51, 9, 15, and 16, respectively (see Table 1 for initial number of variables). The number of variables in the remaining data blocks were unchanged, as differences across samples within each block were significant for all the variables (one-way ANOVA, $p<0.05$). After removal of non-significant variables all data blocks were analysed using PCA and systematic variation was investigated using scores and loadings plots (not shown). From this approach outliers were not detected in any of the blocks analysed.

3.2 Global model of multiple Y-variables using SO-PLS2

Data was initially analysed with PLS2 (see point IV in the data check and preparation described above) prior to SO-PLS2 modelling. Individual PLS2 models were determined for every X-Y block combination and found that two out of twelve individual X-blocks gave low predictive power below 10% validated explained variance (4.1% for X_{03} and a collapsed model with negative validated explained variance for X_{12}). The remaining ten X-blocks were therefore used for subsequent SO-PLS2 modelling.

To determine how the chemical measures from the series of ten X-blocks related to the Y-block (descriptive sensory analysis), the data was modelled using SO-PLS2. With so many X-blocks at hand and no intuitive ordering thereof, one could have computed a vast

number of SO-PLS models, considering that models could be based on: (I) different combinations of only two X-blocks up to as many as ten X-blocks and (II) different order of X-blocks. The simplest SO-PLS model would consist of only the Y-block and any two X-blocks in any order, i.e. models with both X_{01} and X_{02}, X_{01} and X_{04}, X_{01} and X_{05} and so on (note that block X_{03} was left out due to low predictive power). This alone results in $\frac{n!}{(n-r)!} = \frac{10!}{(10-2)!} = 90$ different SO-PLS models with r = 2 X-blocks out of n = 10 X-blocks to choose from. Addition of blocks, so r = 3 X-blocks or 4 X-blocks, rapidly increases the number of models to additional 720 and 5040, respectively. This clearly illustrates that the vast number of models to be computed would go beyond the practicalities of time and computational power, as well as in addition an enormous chance of overfitting. Given this situation, the strategy of progressive modeling approach was used for selection and ordering of the X-blocks to find a more robust SO-PLS model to describe the variation in the Y-block, in an efficient manner.

The most optimal 2 block SO-PLS2 model consisted of blocks X_{06} (CIELab colour measures using 1 component) followed by orthogonalisation of X_{02} (amino acids using 2 components) with respect to the one component from X_{06} (Fig 1.). This resulted in an RMSECV of 1.103, and calibrated and validated explained variances of 59.1 % and 43.2 %, respectively. Additional orthogonalisation of a third X block to compute 3 block SO-PLS2 models did not improve the model further based on RMSECV, number of components, and increased validated explained variance criteria, thus data will only be interpreted up to 2 block SO-PLS2.

The CV-ANOVA gave *p*-values equal to 0.06 and 0.17 for the inclusion of X06 and X02, respectively. This may indicate that although there is a clear improvement in RMSECV in both cases they are not strictly significant. Not knowing the power of the CV-ANOVA test, this result only tells us that one should be careful and not over-interpret the findings as reported below.

The projection of scores from the most optimal 2 block SO-PLS2 model can be visualized on the principal components of prediction (PCP) plots (Figure 2). The explained variance in the PCP plots show that the first and second PCs accounted for 75.2 % and 18.8 % of the variation, respectively, in Y-block that was predicted from the 2 block SO-PLS2 model with X_{06} and X_{02} using 1 and 2 components, respectively.

279 The projection of scores showed relatively close groupings of samples by harvest
280 origin; RVL samples were mostly grouped in the area of both negative PC 1 and 2.
281 Specifically, the eight RVL samples, the EV samples and BV2 projected negatively along PC
282 1 had little association with the majority of the Y loadings (Fig 2). The CWA, LC, McL, BV1,
283 and CV1 samples were projected on positive PC 1, and were characterized by higher values
284 for many of the sensory attribute loadings, as well as three variables from X_{06} from colour
285 measures; hue angle, Chroma, A* (Chroma and A* overlapped with each other), and amino
286 acids from X_{02}; methionine, isoleucine and γ-aminobutyric acid (GABA) (Fig 2). Although it
287 seems that the RVL samples were negatively projected on PC 1 of the scores plot with many
288 of the Y and X loadings, the samples had high values of two variables from the CIELab
289 measures (L* and b* that denote for lightness and yellowness, respectively). The RVL region
290 is known for their hot weather and high growing degree days, which can hinder production of
291 anthocyanin and phenolic compounds. The consequence is an influence on wine colour
292 (Ojeda, Andary, Kraeva, Carbonneau, & Deloire, 2002), which was also reflected in the 2
293 block SO-PLS2 model. The amino acids are known to contribute to wine flavour indirectly
294 by their metabolism by microorganisms and hence the resultant secondary metabolites
295 (Styger, Prior, & Bauer, 2011a, 2011b). These results show that the SO-PLS method provides
296 a strategy to select the X blocks used for modeling and some important ways of visualizing
297 the results.

298 *3.3 Modeling of single Y-variables using SO-PLS1*

299 The SO-PLS2 was taken to further detailed models for single Y-variables using SO-
300 PLS1 to determine the chemical data blocks that explain the differences in specific wine
301 sensory attributes, particularly those that are anecdotally known to be important for wine
302 quality. The SO-PLS1 procedure in block selection was performed in the same manner as
303 described above for SO-PLS2 models. Optimal combinations of data blocks and number of
304 components for each block were first determined for each single Y-variable with PLS1 prior
305 to SO-PLS1. Comparative model parameters were determined for the number of components,
306 RMSECV, and explained variances for calibration and validation (Table 2). Overall, all
307 models were determined with up to four components per data block when modelled with SO-
308 PLS1, the same limit that was set as SO-PLS2 modelling to prevent overfitting of the data. To
309 place the focus on SO-PLS1 applicability, only the sensory attributes (Y-variables) modelled
310 with two or more X-blocks are reported. Four Y-variables were modelled with only one X
311 block (PLS1), which were attributes hue, dark fruit and savoury aromas, and alcohol

mouthfeel. Orthogonal addition of a second X-block did not improve the SO-PLS1 models for these three attributes, therefore these models are not reported. As such, a total of 12 Y-variables were modelled using SO-PLS1: these included sensory attributes (number in brackets) associated with colour (1), aroma (2), taste (1), flavour (4), mouthfeel (3), and aftertaste (1) (Table 2).

The Y-variable modelled with the highest validated explained variance was savoury flavour (F_Savoury, Table 2) (a negative contributor to red wine quality) (Johnson, Hasted, Ristic, & Bastian, 2013). This attribute was modelled with both bound (X_{05}) and non-targeted volatile (X_{04}) compound measures, suggesting that the perception of this sensory attribute was driven by grape-derived volatile compounds. The RMSECV values obtained with SO-PLS1 were consistently lower than the SO-PLS2 model (with the exception of depth of colour), indicating that SO-PLS1 was more effective in describing the systematic variance in the single Y-variables. This is because SO-PLS2, must compromise to fit X blocks to a suite of Y-variables, whereas SO-PLS1 finds the optimal fit of X blocks to only one Y-variable. There are no official cut-off limits with RMSECV values, meaning that the researcher must choose what is acceptable based on the context of the data. The general rule, however, is that lower RMSECV values are more desirable because they denote lower error margins related to the means of the original input data, leading to better prediction accuracy. This must however be balanced with the relative increases in explained variances from the addition of more components and should there be little reduction in RMSECV, it is advisable to use simpler and robust models with lower components.

The model for savoury flavour was interpreted by plotting separately the two PLS models from the two steps in SO-PLS1 (Fig 3). Plotting of the first X block showed regional separation of samples, a useful piece of information to demonstrate the chemical differences by provenance that impact their sensory perception in wines. In particular, the samples from LC and CWA were distinguished by higher relative intensities of both savoury flavours and bound volatile compounds (2,6-dimethoxyphenol, an actinidole, 4-vinylphenol, guaiacol, methyl vanillate, and benzyl alcohol, Fig 3A). The orthogonalised second X block comprising non-targeted volatiles captured additional explained variance. In this instance, McL2, CV2, and CWA2 were projected in the same direction as savoury flavour, along with β-damascenone (a potent grape-derived volatile that enhances fruity aromas and suppresses herbaceous ones), (Pineau, Barbe, Van Leeuwen, & Dubourdieu, 2007) benzaldehyde and to a certain extent, 1-butanol. Meanwhile, RVL3 and RVL6–8, and LC2 had lower levels of

savoury flavour, along with low concentrations of benzaldehyde and β-damascenone (Fig 3B). These samples had higher concentrations of (*Z*)-2-penten-1-ol, 2-methylbutanal, 3-methylbutanal, and benzeneacetaldehyde. The advantages of SO-PLS1, specific Y-variables can be predicted with multiple X-blocks that are orthogonalised with each other.

The 2 block SO-PLS1 models were extended to 3 block SO-PLS1 modeling. Further variance could not be captured beyond two X blocks with meaningful improvement in validated explained variance. This was seen as either improvement being minor in validated explained variance or requirement of a high number of components, leading to over-fitting. Therefore, at this point the modeling procedure was ended.

3.4 Prospective of SO-PLS method in metabolomics research and considerations

The SO-PLS2 modeling was able to incorporate two blocks from a total of 12 blocks. The remaining ten X-blocks did not contain further additional systematic variation that would lead to higher explained variances, and thus more comprehensive models. It may be that modeling was limited because of several factors; the large biological transformation in the sample matrix between grapes (X data) and wine (Y data), the nature of the measurements (with X data being chemical and Y data being perceptual, elaborated below), and perhaps that there were unknown relevant metabolites that were not measured. In spite of this, the results reveal the great potential of the multi-block data analysis approach, in this case by using diverse grape compositional data sets (determined instrumentally) to predict wine sensory properties (as perceived by humans), thereby having profound implications for pre-determining wine sensory characteristics (thus quality and style attributes) “in the vineyard”. The SO-PLS approach could conceivably apply to other research fields, whether food and beverage or biomedical. The current study has shown that it is possible to use up to two X-blocks in SO-PLS1 models to describe a large part of the variation in single Y variables using progressive model optimization, a conservative modelling approach that reduces the chances of over-fitting.

The modeling approach taken with SO-PLS in the current study used a maximum of four components for one X data block. More components can be used, thereby potentially increase the validated explained variance in Y while further reducing the RMSECV, but with a caveat of being wary of over-fitting the model. It is thus important for the data analyst to choose the appropriate number of components suitable for the data type used in the model. Unlike spectroscopic data, where the number of components can be high (Næs, Tomic,

Afseth, Segtnan, & Måge, 2013), the inherent noise associated with sensory measurements (unavoidable inter-panellist variation) calls for a conservative approach that uses fewer components in the models. The exact number may depend on the data type, however the maximum number of components used in this study is most likely suitable to model other sensory data, depending on the degree of increase in explained variance of the model with each component.

Overall, the foremost advantage of SO-PLS is the ability to systematically select X data blocks for analysis of the RMSECV decrease as a function of component combinations, particularly when using many X data blocks. This allows for the selection of the model with the largest increases in explained variances. SO-PLS1 provided substantially increased explained variances for some of the attributes compared to SO-PLS2, which most likely stems from the enhanced ability to match the best fitting X data blocks to each single Y-variables rather than to a whole Y data set. In our case, future steps involve applying the SO-PLS1 method to multiple grape data sets to explore their correlations with sensory perception of wines across multiple vintages, and to determine the consistency in the contribution of grape measures to the modeling of wine sensory attributes. The SO-PLS method has great potential for application in any field that requires prediction of Y data from multiple X blocks, irrespective of research field.

Several limitations are worth considering from the current study. The first is in measurements used for modeling and may include collection of the chemical data, although a good degree of accuracy can be expected from modern analytical instrumentation. It is mainly the perceptual data, having inherent variation due to the nature of using human assessors, such that a certain margin of error in the models is unavoidable thus making predictions extremely challenging. Secondly the choice of the data blocks used in the modeling requires scrutiny from the data analyst to decide whether the optimal X-blocks in the models make sense in the context of the research field. For example, the attribute astringency was best predicted by X_{06} and $X_{08,}$ which is sensible given that these were data sets for colour and anthocyanins, respectively. Pigments contributing to red wine colour (including anthocyanins) are among a range of polyphenolic compounds extracted from red grapes during winemaking that are known to contribute directly to astringency (Brossaud, Cheynier, & Noble, 2001). Should unrelated data blocks give optimal models for astringency, say X_{04} and X_{05} (non-targeted and bound volatile compounds, respectively), a direct relationship would be difficult to explain and likely be correlative than causative. Making

these judgements for models of attributes that have not previously been related to chemical predictors on the other hand will be challenging. Lastly, despite the samples being vinified identically, there will always be unavoidable variations arising from the vinification procedure. This will inevitably be captured as unexplained variance and reflected in the SO-PLS models.

It must be emphasized that the methodology (and results) presented here can be prone to overfitting or over-optimism due to a number of reasons, the most important being the relatively few samples available for building the calibration model as compared to the number of variables and choices/selections that are made. Another reason could be that there is a certain tendency of grouping according to wine region and therefore, all sub-models in the full cross-validation are tested on samples, which are similar to at least one sample in the training set. The fact that the least predictive Y-variables are eliminated, could also have a slight impact. All this means that the prediction results reported should be validated by new data. This also holds for the interpretation. The conclusion of this is that the present study should be considered a feasibility study with some clear indication of how estimation and model fitting can be done and what types of results that can be obtained.

4.0 Conclusions

In the age of big data and using the power of metabolomics, improved methods for modeling diverse datasets and complex phenomenon are still required to reveal underlying relationships that can be overlooked with typical modeling approaches. Thus SO-PLS methodologies were investigated to link grape compositional measures with wine sensory traits determined by human sensory assessment. Modeling of the data with SO-PLS2 showed overall that two X-blocks could be modelled to fit the entire sensory profile of the wines. Further modeling of single Y-variables using SO-PLS1 resulted in lower cross validation error and higher explained variances. Conducting SO-PLS1 with X-data blocks orthogonalised to maintain their data integrity enhanced the modeling of sensory data for single Y-variables. SO-PLS1 was able to determine components that were optimal for each X-data block, which together led to models that better represented the data with higher explained variances than SO-PLS2. The use of SO-PLS provides a strategy for researchers to tackle the issue of analysing and screening multiple data sets to achieve optimal modeling with only important data blocks. The present work has demonstrated the value of the SO-PLS

analysis method in wine analysis and it is expected that this data analysis approach would greatly assist in the advancement of metabolomics research more generally.

Acknowledgements

The authors would like to thank the industry partners CCW Co-operative Ltd, Yalumba Wine Company and Treasury Wine Estates for generously allowing access to vineyards and grape samples. The study was funded by Australia's grape growers and wine makers through their investment body Wine Australia with matching funding from the Australian Federal Government (CSP1201). Sandra Olarte-Mantilla and Trent Johnson are acknowledged for assistance in data collection, and Sue Maffei and Emily Nicholson are acknowledged for their assistance in collecting the grape samples and chemical analyses and the WIC Winemaking service is thanked for producing the small scale wines.

References

Bonnet, J.-L., & Croljzet, J. (1977). Lipoxygenase from tomato fruit: partial purification and study of some properties. *Journal of Food Science, 42*(3), 625-628.

Boss, P., Pearce, A., Zhao, Y., Nicholson, E., Dennis, E., & Jeffery, D. (2015). Potential Grape-Derived Contributions to Volatile Ester Concentrations in Wine. *Molecules, 20*(5), 7845.

Böttcher, C., Boss, P. K., & Davies, C. (2012). Delaying Riesling grape berry ripening with a synthetic auxin affects malic acid metabolism and sugar accumulation, and alters wine sensory characters. *Functional Plant Biology, 39*(9), 745-753.

Brossaud, F., Cheynier, V., & Noble, A. C. (2001). Bitterness and astringency of grape and wine polyphenols. *Australian Journal of Grape and Wine Research, 7*(1), 33-39.

Calderon-Orellana, A., Mercenaro, L., Shackel, K. A., Willits, N., & Matthews, M. A. (2014). Responses of Fruit Uniformity to Deficit Irrigation and Cluster Thinning in Commercial Winegrape Production. *American Journal of Enology and Viticulture*.

Downey, M. O., & Rochfort, S. (2008). Simultaneous separation by reversed-phase high-performance liquid chromatography and mass spectral identification of anthocyanins and flavonols in Shiraz grape skin. *Journal of Chromatography A, 1201*(1), 43-47.

Feron, G., Ayed, C., Qannari, E. M., Courcoux, P., Laboure, H., & Guichard, E. (2014). Understanding aroma release from model cheeses by a statistical multiblock approach on oral processing. *PloS One, 9*(4).

Hanlin, R. L., & Downey, M. O. (2009). Condensed Tannin Accumulation and Composition in Skin of Shiraz and Cabernet Sauvignon Grapes during Berry Development. *American Journal of Enology and Viticulture, 60*(1), 13-23.

Iland, P. G., Bruner, N., Edwards, G., Caloghiris, S., & Willkes, E. (2013). *Chemical analysis of grapes and wine: Techniques and concepts*. Campbelltown, SA: Patrick Iland Wine Promotions.

Indahl, U. G., & Naes, T. (1998). Evaluation of alternative spectral feature extraction methods of textural images for multivariate modelling. *Journal of Chemometrics, 12*(4), 261-278.

Johnson, C. H., Ivanisevic, J., Benton, H. P., & Siuzdak, G. (2015). Bioinformatics: The Next Frontier of Metabolomics. *Analytical Chemistry, 87*(1), 147-156.

Johnson, T. E., Hasted, A., Ristic, R., & Bastian, S. E. P. (2013). Multidimensional scaling (MDS), cluster and descriptive analyses provide preliminary insights into australian Shiraz wine regional characteristics. *Food Quality and Preference, 29*(2), 174-185.

Kalua, C. M., & Boss, P. K. (2009). Evolution of Volatile Compounds during the Development of Cabernet Sauvignon Grapes (Vitis vinifera L.). *Journal of Agricultural and Food Chemistry, 57*(9), 3818-3830.

Langsrud, Ø., & Næs, T. (2003). Optimised score plot by principal components of predictions. *Chemometrics and Intelligent Laboratory Systems, 68*(1–2), 61-74.

Måge, I., Menichelli, E., & Næs, T. (2012). Preference mapping by PO-PLS: Separating common and unique information in several data blocks. *Food Quality and Preference, 24*(1), 8-16.

McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, vol. 445 (pp. 51-56).

Menichelli, E., Almoy, T., Tomic, O., Olsen, N. V., & Naes, T. (2014). SO-PLS as an exploratory tool for path modelling. *Food Quality and Preference, 36*, 122-134.

Mercurio, M. D., Dambergs, R. G., Herderich, M. J., & Smith, P. A. (2007). High Throughput Analysis of Red Wine and Grape Phenolics - Adaptation and Validation of Methyl Cellulose Precipitable Tannin Assay and Modified Somers Color Assay to a Rapid 96 Well Plate Format. *Journal of Agricultural and Food Chemistry, 55*(12), 4651-4657.

Næs, T., Tomic, O., Afseth, N. K., Segtnan, V., & Måge, I. (2013). Multi-block regression based on combinations of orthogonalisation, PLS-regression and canonical correlation analysis. *Chemometrics and Intelligent Laboratory Systems, 124*, 32-42.

Næs, T., Tomic, O., Mevik, B. H., & Martens, H. (2011). Path modelling by sequential PLS regression. *Journal of Chemometrics, 25*(1), 28-40.

Niimi, J., Boss, P. K., Jeffery, D., & Bastian, S. E. P. (2017). Linking the sensory properties and chemical composition of *Vitis vinifera* cv. Cabernet Sauvignon grape berries to wine. *American Journal of Enology and Viticulture, 68*(3), 357-368.

Ojeda, H., Andary, C., Kraeva, E., Carbonneau, A., & Deloire, A. (2002). Influence of pre- and postveraison water deficit on synthesis and concentration of skin phenolic compounds during berry growth of *Vitis vinifera* cv. Shiraz. *American Journal of Enology and Viticulture, 53*(4), 261-267.

Oliphant, T. E. (2007). Python for Scientific Computing. *Computing in Science & Engineering, 9*(3), 10-20.

Peréz, F., & Granger, B. E. (2007). IPython: A System for Interactive Scientific Computing. *Computing in Science & Engineering, 9*(3), 21-29.

Pineau, B., Barbe, J.-C., Van Leeuwen, C., & Dubourdieu, D. (2007). Which Impact for β-Damascenone on Red Wines Aroma? *Journal of Agricultural and Food Chemistry, 55*(10), 4103-4108.

Sarneckis, C. J., Dambergs, R. G., Jones, P., Mercurio, M., Herderich, M. J., & Smith, P. A. (2006). Quantification of condensed tannins by precipitation with methyl cellulose: development and validation of an optimised tool for grape and wine analysis. *Australian Journal of Grape and Wine Research, 12*(1), 39-49.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, (pp. 57-61).

Styger, G., Prior, B., & Bauer, F. F. (2011a). Wine flavor and aroma. *Journal of Industrial Microbiology and Biotechnology, 38*(9), 1145-1159.

Styger, G., Prior, B., & Bauer, F. F. (2011b). Wine flavor and aroma. *Journal of Industrial Microbiology & Biotechnology, 38*(9), 1145.

Tesnière, C., & Verriès, C. (2000). Molecular cloning and expression of cDNAs encoding alcohol dehydrogenases from Vitis vinifera L. during berry development. *Plant Science, 157*(1), 77-88.

Vick, B. A. (1991). A spectrophotometric assay for hydroperoxide lyase. *Lipids, 26*(4), 315-320.

Westerhuis, J. A., Kourti, T., & MacGregor, J. F. (1998). Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics, 12*(5), 301-321.

Wishart, D. S. (2008). Metabolomics: applications to food science and nutrition research. *Trends in Food Science & Technology*, *19*(9), 482-493.

Table 1. Data blocks assigned for data analysis based on the measurements performed and analysis method used for each Cabernet Sauvignon grape sample (n = 25) arising from different vineyards and regions (with references and cross-reference to Supporting Information for additional details of the methods, where applicable).

Data block*	Measurement	Data Dimensions†	Analysis method	Method Reference and Supplementary Experimental Section
X_{01}	Harvest measures	25 × 6	Weight, TSS‡, pH, TA#	(Böttcher, Boss, & Davies, 2012), S-1
X_{02}	Amino acids	75 × 25	HPLC	(Boss, Pearce, Zhao, Nicholson, Dennis, & Jeffery, 2015)
X_{03}	Targeted volatile compounds	75 × 12	GC-MS	S-2
X_{04}	Non-targeted volatile compounds	75 × 27	GC-MS	(Kalua & Boss, 2009)
X_{05}	Bound volatile compounds	75 × 62	GC-MS	S-3
X_{06}	Color	75 × 5	CIELab tristimulus	S-4
X_{07}	Total phenolics and tannins	75 × 3	UV spectrophotometry	(Iland, Bruner, Edwards, Caloghiris, & Willkes, 2013; Mercurio, Dambergs, Herderich, & Smith, 2007; Sarneckis, Dambergs, Jones, Mercurio, Herderich, & Smith, 2006)
X_{08}	Anthocyanins	75 × 11	HPLC	(Downey & Rochfort, 2008)
X_{09}	Tannins	75 × 11	HPLC	(Hanlin & Downey, 2009)
X_{10}	Flavonols	75 × 7	HPLC	(Downey & Rochfort, 2008)
X_{11}	Fatty acids	75 × 31	GC-MS	S-5
X_{12}	Lipoxygenase pathway enzyme activities	75 × 3	Spectrophotometric	(Bonnet & Croljzet, 1977; Tesnière & Verriès, 2000; Vick, 1991)
Y	Sensory profiles	75 × 28	Descriptive analysis	(Niimi, Boss, Jeffery, &

*X block measurements were made on grapes and the Y block measurement was made on wines.
† X01 consists of 25 samples as the inputs were averaged prior to analysis, whereas the remaining X blocks included triplicates of 25 samples, giving a total of 75.
‡ TSS is total soluble solids.
# Titratable acidity.

Table 2. Most optimal 2X block SO-PLS1 models determined for Y-variables (sensory attributes pertaining to wine quality).

Y-Variable[a]	1st X	2nd X	Comp[b]	RMSECV	Cal (%)[c]	Val (%)[d]
F_Savory	X_{05}	X_{04}	3_2	0.492	95.0	69.1
C_Depth	X_{07}	X_{10}	2_2	1.920	78.8	68.6
A_Overall	X_{10}	X_{06}	2_4	0.433	81.5	66.3
AT_Phenolic Length	X_{06}	X_{02}	1_3	0.551	81.6	64.6
MF_Tannin quality	X_{06}	X_{09}	1_3	0.647	77.5	64.1
T_Acid	X_{04}	X_{11}	3_4	0.475	94.5	63.4
MF_Astringency	X_{06}	X_{08}	1_1	0.938	64.9	56.5
A_Dried fruit	X_{06}	X_{02}	1_3	0.714	74.2	51.1
F_Pepper	X_{01}	X_{02}	2_2	0.562	76.3	50.1
F_Dried fruit	X_{10}	X_{06}	2_3	0.707	69.1	50.1
F_Dark fruit	X_{06}	X_{11}	1_1	0.948	58.6	44.2
MF_Body	X_{07}	X_{02}	2_2	0.661	65.9	42.8

a C_ - color, A_ - aroma, T_ - taste, F_ - flavor, MF_ - mouthfeel, AT_ - aftertaste.
b Component values in SO-PLS1 (e.g. 3_4) denotes for number of components in first and second block, respectively
c Cal – calibrated explained variance
d Val – validated explained variance

Figure(s)

Click here to download Figure(s): Figures.docx

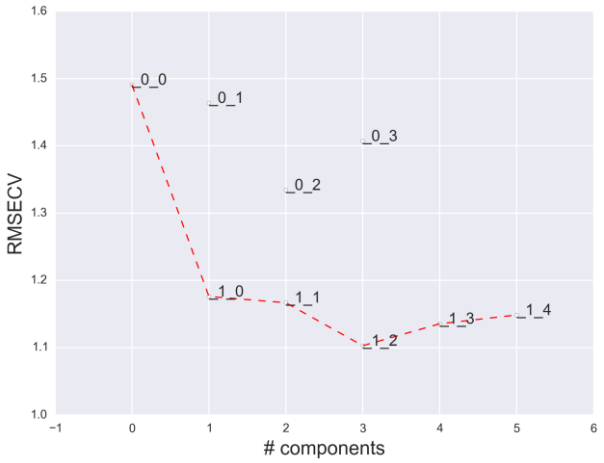

Fig 1. Måge plot showing reduction in RMSECV as a function of total number of model components for 2 block SO-PLS2 with X_{06} and X_{02}. The numbers above points with underscores denote for the number of components for each data set, i.e., components for 1st and 2nd data blocks. In this particular case, using one component for X_{06} and two components for X_{02} (hence _1_2) led to the lowest RMSECV.

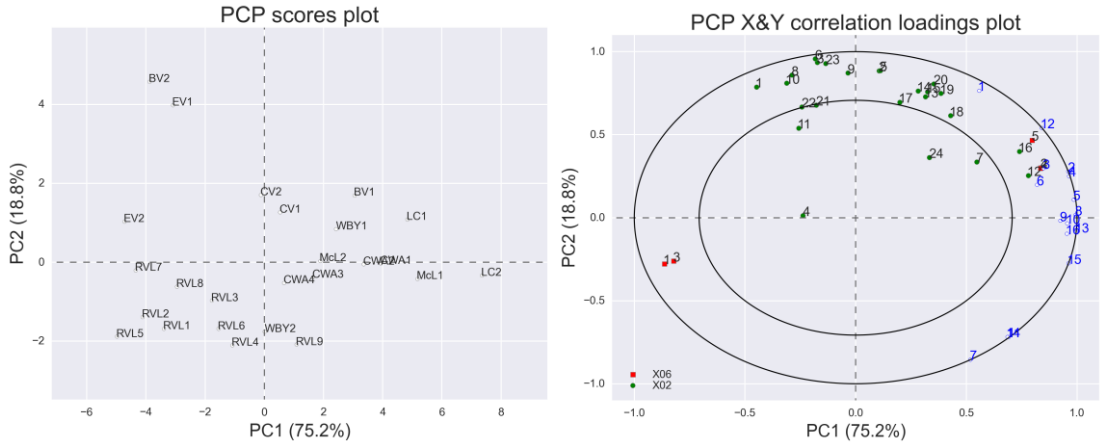

Fig 2. The PCP scores plot (left) and the X and Y correlation loadings plot (right) on the first two principal components, obtained from 2 block SO-PLS2 modeling of Y-block with X_{06} (CIELab) followed by X_{02} (amino acids) chemical data blocks. Variables in blue on the correlation loadings plot denote for Y loadings (sensory attributes) and those in green and red denote for X loadings belonging to X_{06} and X_{02} blocks, respectively. Numbers corresponding to each X and Y loadings are provided in supplementary information (Table S-4). The outer and inner ellipse on the correlations loadings plot indicate 100 % and 50 % of explained variance, respectively. Sample symbols denote for the following: BV-Barossa Valley, CV-Clare Valley, CWA-Coonawarra, EV-Eden Valley, LC-Langhorne Creek, McL-McLaren Vale, RVL-Riverland, and WBY-Wrattonbully.

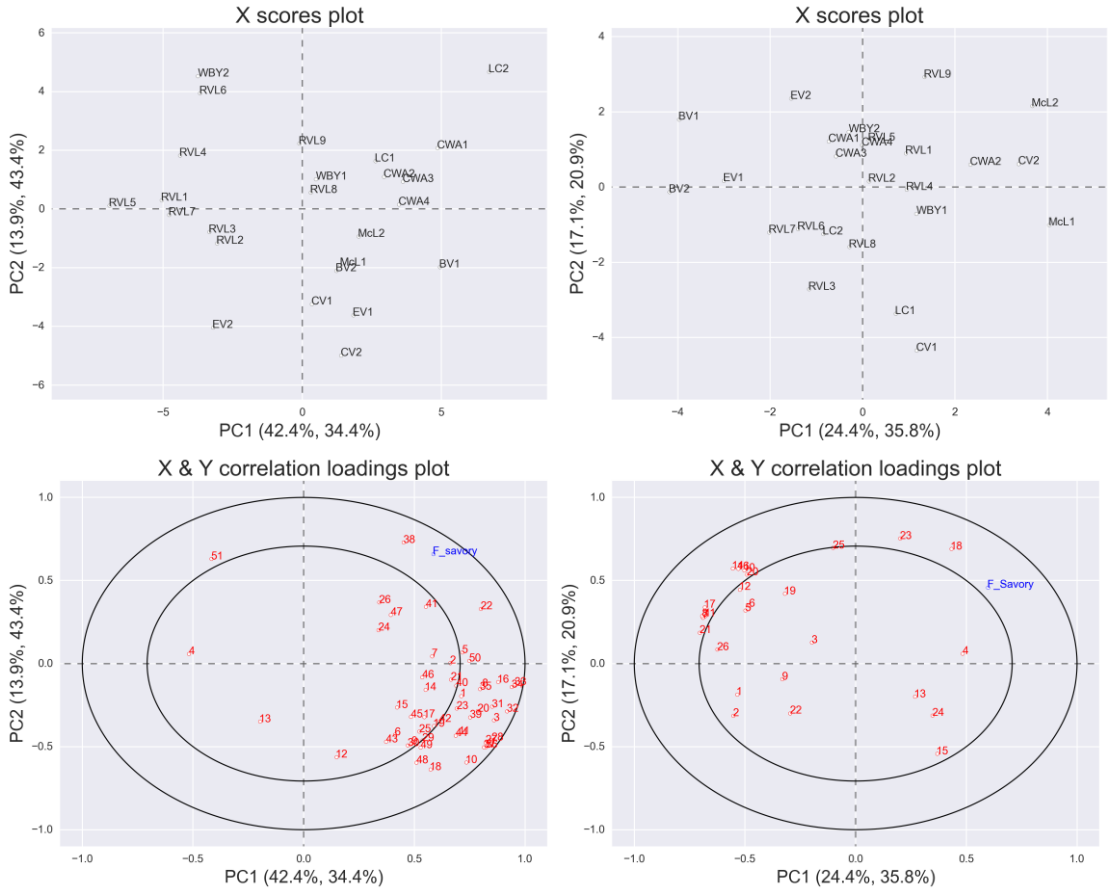

Fig 3. PLS scores and loadings plots for savory flavor attribute using 2 block SO-PLS1, showing (A) scores (top) and loadings (bottom) plots model with the first block X_{05} (left plots) (bound volatiles, S-7), and the second block X_{04} (non-targeted volatiles, Table S-5) after orthogonalization with respect to the 3 components from X_{05} (right plots). The outer and inner ellipse on the correlations loadings plot indicate 100% and 50% of explained variance, respectively. Sample symbols denote for the following: BV-Barossa Valley, CV-Clare Valley, CWA-Coonawarra, EV-Eden Valley, LC-Langhorne Creek, McL-McLaren Vale, RVL-Riverland, and WBY-Wrattonbully.

Supplementary Material

Click here to download Supplementary Material: Supplementary Material.docx