Talanta 224 (2021) 121908

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Short communication

Identifying key wavenumbers that improve prediction of amylose in rice samples utilizing advanced wavenumber selection techniques

Puneet Mishra [a,*], Ernst J. Woltering [a,b]

[a] Wageningen Food and Biobased Research, Bornse Weilanden 9, P.O. Box 17, 6700AA, Wageningen, the Netherlands
[b] Horticulture and Product Physiology Group, Wageningen University, Droevendaalsesteeg 1, P.O. Box 630, 6700AP, Wageningen, the Netherlands

ARTICLE INFO

Keywords:
Chemometrics
Feature selection
Multi-spectral
Food chemistry

ABSTRACT

This study utilizes advanced wavenumber selection techniques to improve the prediction of amylose content in grounded rice samples with near-infrared spectroscopy. Four different wavenumber selection techniques, i.e. covariate selection (CovSel), variable combination population analysis (VCPA), bootstrapping soft shrinkage (BOSS) and variable combination population analysis-iteratively retains informative variables (VCPA-IRIV), were used for model optimization and key wavenumbers selection. The results of the several wavenumber selection techniques were compared with the predictions reported previously on the same data set. All the four wavenumber selection techniques improved the predictive performance of amylose in rice samples. The best performance was obtained with VCPA, where, with only 11 wavenumbers-based model, the prediction error was reduced by 19% compared to what reported previously on the same data set. The selected wavenumbers can help in development of low-cost multi-spectral sensors for amylose prediction in rice samples.

Credit

Puneet Mishra: Conceptualization, Data curation, Investigation, Writing, Ernst Woltering: Writing – review & editing.

1. Introduction

Amylose concentration in rice is the key quality attribute related to its eating quality [1]. Amylose content is correlated with the retrogradation behavior, influencing the textural properties of cooked rice and the viscoelasticity dynamics of rice starch gel. Amylose content is an important biomarker for screening rice genotypes in breeding programs [2]. The amylose content in rice can range from 0 to even >26% [3]. High amylose rice is gaining huge attention due to its associated health benefits such as slow digestion to glucose, which allows management of health conditions such as diabetes [4]. Traditional methods to determine amylose content includes iodine reaction coupled with potentiometric or amperometric titration [2], differential scanning calorimetry [5] and chromatography [6]. However, a main drawback with traditional wet chemistry approaches is that they are time and labor intensive, and usually have higher complexity related in terms of sample preparation. Furthermore, wet chemistry techniques are not suitable for non-destructive in-line implementation.

In recent years, several applications of near-infrared (NIR) spectroscopy can be found related to rapid nondestructive prediction of chemical components in agri-food products [7]. NIR spectroscopy allows capturing the physical and chemical properties of samples as a function of light scattering and absorption, respectively [8]. In relation to amylose prediction in rice, a previous report demonstrated that with NIR spectroscopy (12,000–4000 cm^{-1}), a prediction R^2 and error of 0.88 and 1.938%, respectively, were achievable [2]. Such high correlation and low errors were obtained by using the interval and window based wavenumber selection techniques popular in the chemometrics modelling [9]. However, the interval and window based wavenumber selection techniques do not allow identifying the discrete bands related to the property of interest [10]. The interval and window based techniques rely on several user defined parameters such as interval or window size, number of maximum latent wavenumbers to model for each interval or window and criterion for selecting the intervals. Furthermore, the selected intervals and windows are insufficient for gaining a better understanding of background chemistry and for development of low cost multi-spectral systems.

In recent years, several new wavenumber selection methods have emerged for use in the analysis of NIR data [11–14]. Of particular

* Corresponding author.
E-mail address: puneet.mishra@wur.nl (P. Mishra).

https://doi.org/10.1016/j.talanta.2020.121908
Received 30 October 2020; Received in revised form 14 November 2020; Accepted 17 November 2020
Available online 25 November 2020

Table 1
A summary of calibration and test data set used in this study. The data has exactly the same samples in the calibration and test set as used in the previous work [2].

Data set	Spectra	Amylose concentration (mean ± std) (%)
Calibration	203 × 643	19.7 ± 5.3
Test	110 × 643	20 ± 5

interest are the techniques that allow selection of discrete sets of wavenumbers. Some of the key techniques are covariate selection (CovSel) [11], variable combination population analysis (VCPA) [13], bootstrapping soft shrinkage (BOSS) [12] and variable combination population analysis-iteratively retains information wavenumbers (VCPA-IRIV) [14]. These recent techniques were not available a few years ago, and therefore, most of the data analysis was limited just to either the PLS regression or its variants, such as interval and window based PLS regression [9], for wavenumber selection, as also used in the recent work related to amylose prediction in rice [2]. Hence, this works aims to highlight and compare four major, recently developed, wavenumber selection techniques for optimizing the NIR models for predicting amylose content in rice samples. It was hypothesized that the new discrete wavenumber selection techniques can improve the model accuracies compared to those previously reported on the same data set using interval and window based wavenumber selection approaches. The four discrete wavenumber selection techniques used were CovSel, VCPA, BOSS and VCPA-IRIV. The key wavenumbers selected by the techniques are discussed and provided in a table, such to facilitate the improvement of NIR data modelling and the development of low cost multi-spectral systems for amylose content prediction in rice.

2. Materials and methods

2.1. Rice data set

The data set used in this study is publicly available as data in brief [15] and is related to scientific publication utilizing NIR spectroscopy for amylose determination in rice [2]. The data set consist of NIR spectra and reference amylose measurement from sixteen rice varieties grown at 4 different locations related to a Portuguese Rice Breeding Program executed along four seasons (2012–2015), providing 168 samples. In addition, samples from 11 rice varieties sourced from International Rice Research Institute, Philippines [2]. The NIR measurement were performed on rice flour obtained by grinding 20g of rice samples in a Cyclone Sample Mill (Falling number 3100, Perten, Sweden). The NIR measurements were performed in transflection mode using the MPA equipment (Bruker Optics, Germany). For each rice sample, 16 successive scans were performed, over a wavenumber range (12,000–4000 cm^{-1}), at 16 cm^{-1} of resolution. The reference amylose content was determined using the standard iodine colorimetric method and more details can be found in Refs. [2].

For our modelling we applied exactly the same calibration and test sets that were used for the modelling in the previous work [2]. The exact same data partition was possible as the labels were provided along with the data set in the data in brief [15]. A further description of calibration and test set is provided in Table 1.

2.2. Data pre-processing

The NIR data range was reduced from 12000 to 3595 cm^{-1} (833–2781 nm) to 8933–3595 cm^{-1} (1119–2781 nm) due to very low absorbance in the wavenumbers range 12000–8932 cm^{-1} (833–1119 nm). To have a fair comparison of different wavenumber selection techniques the same pre-processing i.e. 2nd derivative (Savitzky-Golay [16] window 51 and 2nd order polynomial) was used. The 2nd derivative was used to reveal the underlying peaks to facilitate the wavenumber selection techniques [17]. All data analysis was performed in MATLAB 2018b, MathWorks, Natick, USA.

2.3. Wavenumber selection techniques

2.3.1. Covariate selection

Covariance selection (CovSel) is a popular chemometric technique for selecting discrete wavenumbers [11]. In CovSel, wavenumber selection is accomplished by iterating two steps i.e. the wavenumber having maximum covariance with the response(s) is selected and later both the predictor and the response matrices are orthogonalized with respect to the selected wavenumber. These two steps are repeated until a pre-defined criterion is met. In this study, the venetian blind cross-validation approach was used to identify optimal number of wavenumbers which lead to the minimum root mean square error. CovSel was implemented by means of the MBA-GUI toolbox [18].

2.3.2. Variable combination population analysis

Variable combination population analysis (VCPA) is a two-step procedure [13]. First, an exponentially decreasing function (EDF) is employed to determine the number of wavenumbers to keep and continuously shrink the wavenumber space. Second, in each EDF run, a binary matrix sampling (BMS) strategy that gives each wavenumber the same chance to be selected and generates different wavenumber combinations is used to produce a population of subsets to construct a population of sub-models. Then, model population analysis (MPA) is employed to find the wavenumber subsets with the lowest root mean square error of cross validation (RMSECV). The frequency of each wavenumber appearing in the best 10% of sub-models is computed. The wavenumbers with highest frequency are the most important and vice versa. The VCPA was tested using the free codes found at:

https://nl.mathworks.com/matlabcentral/fileexchange/47739-vcpa-1-1-zip.

2.3.3. Bootstrapping soft shrinkage

Bootstrapping soft shrinkage (BOSS) [12] combines the ideas of weighted bootstrap sampling and model population analysis. The weights of wavenumbers are determined based on the absolute values of the regression coefficients. Weighted bootstrap sampling is applied according to the weights to generate sub-models and model population analysis is used to analyze the sub-models to update weights for wavenumbers. During optimization, soft shrinkage is imposed, in which less important wavenumbers are assigned smaller weights. The algorithm runs iteratively and terminates when the number of wavenumbers reaches one. The optimal wavenumbers carrying low cross-validation error (RMSECV) are retained and a new calibration is established with the retained wavenumbers. BOSS was implemented in MATLAB (2018b, Natick, MA, USA) using the freely available codes available at the website:

http://www.mathworks.com/matlabcentral/fileexchange/52770-boss.

2.3.4. Variable combination population analysis-iteratively retains information wavenumbers

Variable combination population analysis-iteratively retains informative variables (VCPA-IRIV) [14] is a VCPA-based hybrid strategy which continuously shrinks the wavenumber space using VCPA as a first step. It then employs iteratively retaining informative wavenumbers (IRIV) [19] to carry out further optimization in the second step. It takes advantage of VCPA and IRIV, and makes up for each one's drawbacks to deal with high numbers of wavenumbers. VCPA-IRIV was tested using the free code from:

https://nl.mathworks.com/matlabcentral/fileexchange/70232-vcpa-based-hybrid-strategy?s_tid=FX_rc2_behav.

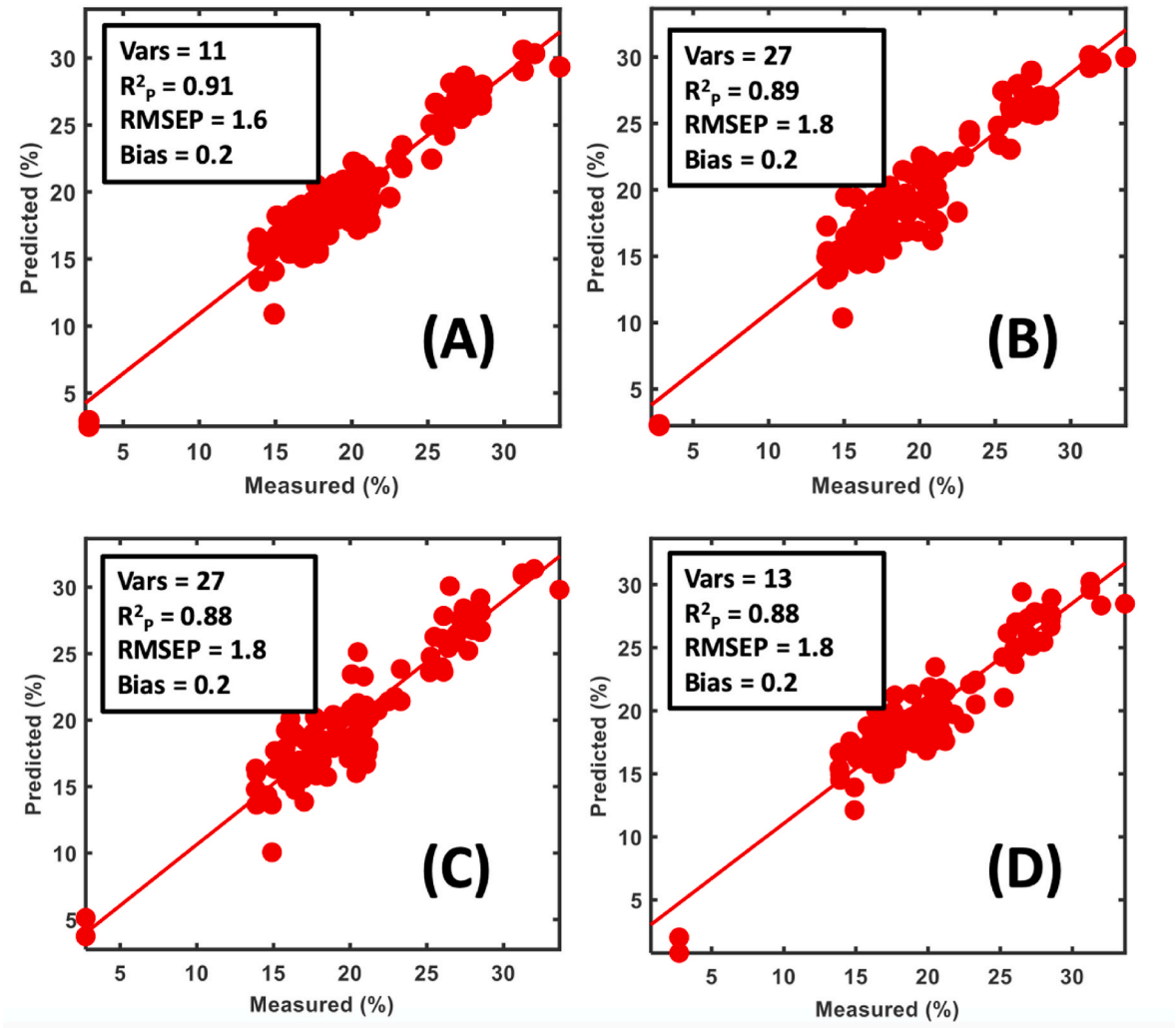

Fig. 1. A summary of models predicting amylose (%) in ground rice samples developed with selected wavenumbers. (A) Variable combination population analysis (VCPA), (B) Bootstrapping soft shrinkage (BOSS), (C) Variable combination population analysis-iteratively retains informative variables (VCPA-IRIV), and (D) covariate selection (CovSel).

Table 2
A summary of selected wavenumbers from each wavenumber selection technique. Variable combination population analysis (VCPA), bootstrapping soft shrinkage (BOSS), variable combination population analysis-iteratively retains informative variables (VCPA-IRIV), and covariate selection (CovSel). The best performing wavenumbers corresponding to VCPA are highlighted in red.

Technique	Selected wavenumbers (cm^{-1})					
	3000–3999	4000–4999	5000–5999	6000–6999	7000–7999	8000–8999
VCPA		4706		6966, 6997	7005,7275, 7360, 7637, 7653, 7707, 7714	8100
BOSS	3927	4690, 4698, 4706, 4752, 4760	5331	6133, 6141, 6966	7259, 7367, 7498, 7529, 7630, 7645, 7699, 7707, 7714, 7722, 7761, 7768	8069, 8100, 8108, 8231, 8262
VCPA-IRIV		4359, 4667, 4675, 4713, 4721	5523	6781, 6966, 6974, 6989	7143, 7151, 7167, 7190, 7197, 7251, 7560, 7622, 7630, 7714, 7768, 7776, 7792, 7830,	8085, 8100, 8154
CovSel				6858, 6989, 6711	7089, 7414, 7522, 7722, 7630	8131, 8308, 8355, 8339, 8270

3. Results

The results of all four wavenumber selection techniques are shown in Fig. 1. All the four wavenumbers selection techniques improved the model performance compared to the model reported previously on the same data set using the interval and window based methods [2]. The best reported performance with the interval and window based methods was $R^2_P = 0.88$ and RMSEP = 1.9% [2]. In the present study, out of the four techniques, the VCPA attained the best performance with $R^2_P = 0.91$ and RMSEP = 1.6%. Improvement by VCPA, lead to a 19% decrease in the prediction error compared to reported in previous work [2]. Further, VCPA attained this by using only 11 wavenumbers (discrete wavenumbers) out of the 643 wavenumbers originally present in the NIR data.

The BOSS approach performed the second best in term of the prediction error but selected 27 wavenumbers which is almost double to that obtained with VCPA. Followed by BOSS, both the VCPA-IRIV and CovSel performed similar in terms of prediction error but CovSel

selected only 13 wavenumbers compared to 27 wavenumbers by VCPA-IRIV. Hence, in terms of lowest number of selected wavenumber and obtaining the lowest prediction error, VCPA performed the best.

A summary of the selected wavenumbers with different techniques is shown in Table 2. For an easy understanding the wavenumber were partitioned into 6 classes i.e. 3000-3999, 4000–4999, 5000–5999, 6000–6999, 7000–7999 and 8000-8999 cm^{-1}. In the previous work, the interval and window based methods identified wavenumber ranges of 8941–8194 cm^{-1}; 5592–5045 cm^{-1}; and 4683–4335 cm^{-1} [2]. However, the discrete wavenumbers selected in this study by VCPA were not related to any of the wavenumber ranges identified previously [2]. The VCPA identified wavenumber 4706 cm^{-1} is related to OH combination bonds related to polysaccharides [20,21]. The wavenumbers 6766 and 6997 cm^{-1} can be assigned to OH 1st overtones of crystalline cellulose and OH groups with H-bonds of intermediate strength [20,21]. The wavenumbers 7005, 7275, 7360, 7637, 7653, 7707 and 7714 cm^{-1} can be assigned to free OH group or weakly bonded OH and 1st overtones of CH, CH_2 and CH_3 [20,21]. The wavenumber 8100 cm^{-1} can be assigned to the 2nd overtones of the CH, CH_2 and CH_3 [20,21].

4. Conclusions

The study showed that the optimization of NIR models with discrete wavenumber selection techniques improved the prediction of NIR models for amylose prediction. Out of the four wavenumber selection techniques used, the VCPA attained the lowest prediction error which was almost 19% lower compared to the prediction error reported previously on the same data set using the interval based wavenumber selection techniques. VCPA selected only 11 wavenumbers out of 693 and were easily assigned to the overtones for OH, CH, CH_2 and CH_3 present in polysaccharides such as amylose. The selected wavenumbers can be used to either improve the already developed models or to build low cost multi-spectral systems for amylose prediction in rice samples. The selected wavenumbers were: 4706, 6966, 6997, 7005, 7275, 7360, 7637, 7653, 7707, 7714 and 8100 cm^{-1}.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Y. Sun, G. Jiao, Z. Liu, X. Zhang, J. Li, X. Guo, W. Du, J. Du, F. Francis, Y. Zhao, L. Xia, Generation of high-amylose rice through CRISPR/Cas9-Mediated targeted mutagenesis of starch branching enzymes, Front. Plant Sci. 8 (2017) 298.

[2] P.S. Sampaio, A. Soares, A. Castanho, A.S. Almeida, J. Oliveira, C. Brites, Optimization of rice amylose determination by NIR-spectroscopy using PLS chemometrics algorithms, Food Chem. 242 (2018) 196–204.

[3] B.O. Juliano, C.M. Perez, A.B. Blakeney, T. Castillo, N. Kongseree, B. Laignelet, E. T. Lapis, V.V.S. Murty, C.M. Paule, B.D. Webb, International cooperative testing on the amylose content of milled rice, Starch - Stärke 33 (1981) 157–162.

[4] K. Tao, W. Yu, S. Prakash, R.G. Gilbert, High-amylose rice: starch molecular structural features controlling cooked rice texture and preference, Carbohydr. Polym. 219 (2019) 251–260.

[5] D. Sievert, J. Holm, Determination of amylose by differential scanning calorimetry, Starch - Stärke 45 (1993) 136–139.

[6] J.M. Franco, M.A. Murado, M.I.G. Siso, J. Miron, M.P. Gonzalez, A HPLC method for specific determination ofα-amylase and glucoamylase in complex enzymatic preparations, Chromatographia 27 (1989) 328–332.

[7] J.U. Porep, D.R. Kammerer, R. Carle, On-line application of near infrared (NIR) spectroscopy in food production, Trends Food Sci. Technol. 46 (2015) 211–230.

[8] C. Pasquini, Near infrared spectroscopy: a mature analytical technique with new perspectives – a review, Anal. Chim. Acta 1026 (2018) 8–36.

[9] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, Appl. Spectrosc. 54 (2000) 413–419.

[10] P. Mishra, E. Woltering, B. Brouwer, E. Hogeveen-van Echtelt, Improving moisture and soluble solids content prediction in pear fruit using near-infrared spectroscopy with wavenumber selection and model updating approach, Postharvest Biol. Technol. 171 (2021) 111348.

[11] J.M. Roger, B. Palagos, D. Bertrand, E. Fernandez-Ahumada, CovSel: wavenumber selection for highly multivariate and multi-response calibration: application to IR spectroscopy, Chemometr. Intell. Lab. Syst. 106 (2011) 216–223.

[12] B.-C. Deng, Y.-H. Yun, D.-S. Cao, Y.-L. Yin, W.-T. Wang, H.-M. Lu, Q.-Y. Luo, Y.-Z. Liang, A bootstrapping soft shrinkage approach for wavenumber selection in chemical modeling, Anal. Chim. Acta 908 (2016) 63–74.

[13] Y.-H. Yun, W.-T. Wang, B.-C. Deng, G.-B. Lai, X.-b. Liu, D.-B. Ren, Y.-Z. Liang, W. Fan, Q.-S. Xu, Using wavenumber combination population analysis for wavenumber selection in multivariate calibration, Anal. Chim. Acta 862 (2015) 14–23.

[14] Y.-H. Yun, J. Bin, D.-L. Liu, L. Xu, T.-L. Yan, D.-S. Cao, Q.-S. Xu, A hybrid wavenumber selection strategy based on continuous shrinkage of wavenumber space in multivariate calibration, Anal. Chim. Acta 1058 (2019) 58–69.

[15] P. Sampaio, A. Soares, A. Castanho, A.S. Almeida, J. Oliveira, C. Brites, Dataset of Near-infrared spectroscopy measurement for amylose determination using PLS algorithms, Data in Brief 15 (2017) 389–396.

[16] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639.

[17] P. Mishra, A. Biancolillo, J.M. Roger, F. Marini, D.N. Rutledge, New data preprocessing trends based on ensemble of multiple preprocessing techniques, Trac. Trends Anal. Chem. (2020) 116045.

[18] P. Mishra, J.M. Roger, D.N. Rutledge, A. Biancolillo, F. Marini, A. Nordon, D. Jouan-Rimbaud-Bouveresse, MBA-GUI, A Chemometric Graphical User Interface for Multi-Block Data Visualisation, Regression, Classification, Wavenumber Selection and Automated Pre-processing, Chemometrics and Intelligent Laboratory Systems, 2020, p. 104139.

[19] Y.-H. Yun, W.-T. Wang, M.-L. Tan, Y.-Z. Liang, H.-D. Li, D.-S. Cao, H.-M. Lu, Q.-S. Xu, A strategy that iteratively retains informative wavenumbers for selecting optimal wavenumber subset in multivariate calibration, Anal. Chim. Acta 807 (2014) 36–43.

[20] B.G. Osborne, T. Fearn, P.H. Hindle, Practical NIR Spectroscopy with Applications in Food and Beverage Analysis, Longman scientific and technical1993.

[21] B.G. Osborne, Near-Infrared Spectroscopy in Food Analysis, Encyclopedia of Analytical Chemistry, 2006.