JUSSARA VALENTE ROQUE

NEW FEATURES OF ORDERED PREDICTORS SELECTION FOR MULTIVARIATE REGRESSION AND CLASSIFICATION

Thesis submitted to the Universidade Federal de Viçosa, as part of the requirements of the Agrochemistry Graduate Program, for obtaining the *Doctor Scientiae* degree.

Advisor: Reinaldo Francisco Teófilo

Co-Advisor: Luiz Alexandre Peternelli

VIÇOSA – MINAS GERAIS
2019

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
R786n
2019

Roque, Jussara Valente, 1988-
New features of ordered predictors selection for multivariate regression and classification / Jussara Valente Roque. – Viçosa, MG, 2019.
109 f. : il. (algumas color.) ; 29 cm.

Texto em inglês.
Orientador: Reinaldo Francisco Teofilo.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Variáveis (Matemática). 2. Sistemas de reconhecimento de padrões. 3. Calibração. 4. Mínimos quadrados. 5. Análise discriminatória . I. Universidade Federal de Viçosa. Departamento de Química. Programa de Pós-Graduação em Agroquímica. II. Título.

CDD 22. ed. 515.64

JUSSARA VALENTE ROQUE

NEW FEATURES OF ORDERED PREDICTORS SELECTION FOR MULTIVARIATE REGRESSION AND CLASSIFICATION

Thesis submitted to the Universidade Federal de Viçosa, as part of the requirements of the Agrochemistry Graduate Program, for obtaining the *Doctor Scientiae* degree.

APPROVED: August 2nd, 2019.

Jussara Valente Roque
(Author)

Reinaldo Francisco Teófilo
(Advisor)

```
function dedicatoria(leitor)
display('Dedico esta tese a ...')
switch leitor
   case {'mãe','pai'}
      display('meus pais, João e Sônia!')
   case {'irmãos'}
      display('meus irmãos, Marco e Tati!')
   otherwise
      display('todos os amantes da quimiometria!')
end
```

AGRADECIMENTOS

Agradeço aos meus pais João e Sônia, e meus irmãos Marco e Tati, pelo amor, orações, apoio e paciência.

Ao Prof. Reinaldo F. Teófilo pela orientação desde os tempos de mestrado, pela confiança em mim depositada, pela amizade que construímos, e acima de tudo pela paciência e por estar sempre disposto a ensinar!

Ao Prof. Luiz A. Peternelli pelas inúmeras vezes que me recebeu para proveitosas discussões sobre estatística e quimiometria.

Aos amigos do MCDALab, de hoje e de outros tempos, Danilo, Duda, Gilmara, Helder, Igor, Ítalo, João, Larissa, Laura, Mariana, Nathália, Paula, Raquel, Tamires, Ulisses, Wesley e Wilson, que conviveram comigo durante os anos que por lá estive, pela amizade e por estarem sempre dispostos a ajudar. Especial agradecimento à Larissa, Nathália, Wesley e Wilson pela amizade que construímos desde 2017. Foi muito bom compartilhar todas aquelas experiências com vocês! Especial também para a Mariana, que esteve sempre presente nos anos iniciais do meu doutorado, sempre com muita paciência em ouvir e ajudar! E por último, mas não menos importante, agradeço imensamente ao Wilson, por compartilhar comigo a paixão pela quimiometria e me ajudar imensamente na conclusão deste trabalho!

Aos amigos de Viçosa, Armanda, Dayana, Sara e Raphael, Yara e Ronan, e ao Mateus, pela amizade, pelas longas conversas e por todo o apoio durante o doutorado.

Aos professores Marcelo Martins de Sena, João Paulo Ataide Martins, Mariana Ramos de Almeida, e a Camila Assis que gentilmente aceitaram o convite para compor a banca examinadora.

À Juliano S. Ribeiro, Eduardo B. de Melo, Nathália A. Porto, Luiz A. Peternelli, e Hebert V. Pereira por fornecerem alguns conjuntos de dados utilizados nesta tese.

À Universidade Federal de Viçosa, ao Departamento de Química, e ao Programa de Pós-graduação em Agroquímica pela oportunidade de realizar este trabalho.

À FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais) e CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pela concessão da bolsa de estudo. Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo financiamento (Processos 310303/2015-0 e 310503/2015-9).

A todos aqueles que contribuíram direta ou indiretamente para a realização deste trabalho.

A única maneira de fazer um excelente

trabalho é amar o que você faz.

Steve Jobs

ABSTRACT

ROQUE, Jussara Valente, D.Sc., Universidade Federal de Viçosa, August, 2019. **New Features of Ordered Predictors Selection for Multivariate Regression and Classification.** Advisor: Reinaldo Francisco Teófilo. Co-advisor: Luiz Alexandre Peternelli.

New variable selection methods for multivariate regression and classification based on ordered predictors selection (OPS) were developed in this work. Initially, the new OPS strategies for regression were developed and applied to the six datasets used in the original OPS paper to compare their prediction performances. After that, twelve new datasets were used to test and compare the new OPS approaches for regression with other variable selection methods, genetic algorithm (GA), the interval successive projections algorithm for partial least squares (*i*SPA), and recursive weighted partial least squares (rPLS). Simulated datasets were used to evaluate the computational performance of variable selection methods, being then the new OPS approaches for regression, GA, *i*SPA, and rPLS. All methods were evaluated by using a central composite design varying the matrix dimensions of simulated datasets and the number of latent variables. For classification, OPS methods for feature selection in the discriminant analysis (OPSDA) were developed. OPSDA methods were applied to three datasets with different numbers of classes, and classification models were built using different classification methods. The new OPS approaches for regression outperformed the first OPS version and the other variable selection methods. Results showed that in addition to higher predictive capacity, the accuracy in the selection of expected variables is highly superior with the new OPS approaches for regression. The computational performance of OPS approaches was mainly influenced by the number columns of the data matrix, as well as the GA. On the other hand, *i*SPA and rPLS were mainly influenced by the number of rows. In classification, the OPSDA methods provided the best set of selected variables to build more predictive models using different classification methods. Besides, they could be applied to classification problems, independent of the number of classes. Overall, the new OPS methods provided the best set of selected variables to build more predictive and interpretative regression and classification models. The new OPS methods proved to be efficient for variable selection in different types of datasets.

Keywords: Variable Selection, Multivariate Regression, Supervised Pattern Recognition.

RESUMO

ROQUE, Jussara Valente, D.Sc., Universidade Federal de Viçosa, agosto de 2019. **Novas abordagens da seleção dos preditores ordenados para regressão multivariada e classificação.** Orientador: Reinaldo Francisco Teófilo. Coorientador: Luiz Alexandre Peternelli.

Neste trabalho foram desenvolvidos novos métodos de seleção de variáveis para regressão multivariada e classificação baseados na seleção dos preditores ordenados (OPS). Inicialmente, novas estratégias do OPS para regressão foram desenvolvidas e aplicadas nos seis conjuntos de dados usados no artigo original do OPS. Em seguida, doze novos conjuntos de dados foram usados para testar e comparar as novas abordagens do OPS para regressão com outros métodos de seleção de variáveis, como o algoritmo genético (GA), o algoritmo de projeções sucessivas em intervalos para quadrados mínimos parciais (*i*SPA) e quadrados mínimos parciais ponderados recursivos (rPLS). Conjuntos de dados simulados foram usados para avaliar o desempenho computacional de métodos de seleção de variáveis, sendo eles as novas abordagens OPS para regressão, GA, *i*SPA e rPLS. Esta análise foi realizada usando um planejamento composto central variando as dimensões dos dados simulados e o número de variáveis latentes. Para classificação, foram desenvolvidos novos métodos OPS para análise discriminante (OPSDA). Diferentes métodos de classificação foram usados para construir modelos. Os métodos OPSDA foram aplicados em três conjuntos de dados com diferentes números de classes. As novas abordagens do OPS para regressão selecionaram variáveis que proporcionaram a construção de modelos mais preditivos que a primeira versão OPS e que os outros métodos de seleção de variáveis. Os resultados mostraram que, além de maior capacidade preditiva, a exatidão na seleção das variáveis interpretativas foi superior com os novos métodos OPS para regressão. O desempenho computacional desses métodos foi influenciado principalmente pelo número de colunas da matriz de dados, bem como para o GA. Por outro lado, o *i*SPA e rPLS foram influenciados principalmente pelo número de linhas. Os métodos OPSDA forneceram o melhor conjunto de variáveis selecionadas para construir modelos mais preditivos usando diferentes métodos de classificação, independentemente do número de classes. No geral, os novos métodos OPS forneceram o melhor conjunto de variáveis selecionadas para construir modelos de regressão e classificação mais preditivos e interpretativos, provando serem eficientes para seleção de variáveis em diferentes tipos de conjuntos de dados.

Palavras-chave: Seleção de variáveis, Regressão multivariada, Reconhecimento de padrões supervisionado.

LIST OF ABBREVIATIONS

Abbreviation	Definition
autoOPS	Automatic ordered predictors selection
autoOPSDA	Automatic ordered predictors selection for discriminant analysis
BIDIAG	Bidiagonalization algorithm
Cal	Calibration
CARS	Competitive adaptative reweighted sampling
CCD	Central composite design
COR	Correlation between each column of matrix **X** with **y**
COV	Covariance
CV	Cross-validation
ECMLR	Equidistant combination multiple linear regression
ErrorCV	Classification error of cross-validation
feedOPS	Feedback ordered predictors selection
feedOPSDA	Feedback ordered predictors selection for discriminant analysis
FN	False negative
FP	False positive
full-PLS-DA	All variables with PLS-DA
full-SVM-DA	All variables with SVM-DA
GA	Genetic algorithm
GC	Gas chromatography
HCA	Hierarchical cluster analysis
***h*Mod**	Number of latent variables of the model
***h*OPS**	Number of latent variables to generate informative vectors in OPS
iOPS	Interval ordered predictors selection
iOPSDA	Interval ordered predictors selection for discriminant analysis
***i*PLS**	Interval partial least squares
***i*SPA**	Intervals successive projections algorithm for PLS
IVS	Interactive variable selection
***k*-NN**	*k*-nearest neighbor
LDA	Linear discriminant analysis

Abbreviation	Definition
MATLAB	Matrix laboratory
MLR	Multiple linear regression
MMP-2	Matrix metalloproteinases type 2
MS	Mass spectrometry
MSC	Multiplicative scatter correction
NAS	Net analyte signal
NIPALS	Nonlinear iterative partial least squares
NIR	Near-infrared spectroscopy
nlvs	Number of latent variables
NMR	Nuclear magnetic resonance spectroscopy
nvars	Number of selected variables
OPS	Ordered predictors selection
OPSDA	Ordered predictors selection for discriminant analysis
OPSDA*pls*	Selected variables by OPSDA with PLS-DA
OPSDA*svm*	Selected variables by OPSDA with SVM-DA
OPS*v*1	First version of ordered predictors selection
PCA	Principal component analysis
PCR	Principal component regression
PLS	Partial least squares
PLS-DA	Partial least squares for discriminant analysis
Pred	Prediction
PRODALL	The product of all single informative vectors simultaneously
QSAR	Quantitative structure-activity relationship
RBF	Radial basis function
R_c	Correlation coefficient of calibration
R_{cv}	Correlation coefficient of cross-validation
rd	Relative difference
REG	Regression coefficients
RMSEC	Root mean square error of calibration
RMSECV	Root mean square error of cross-validation
RMSEP	Root mean square error of prediction

Abbreviation	Definition
R_p	Correlation coefficient of prediction
rPLS	Recursive weighted partial least squares
sc	Selection criterion
SIMCA	Soft independent modeling of class analogy
SMC	Significance multivariate correlation
SNV	Standard normal variate
SPA	Successive projections algorithm
SPA-MLR	Successive projections algorithm for multiple linear regression
SQR	Residual information of the reconstructed matrix with *h*OPS latent variables
ST-PLS	Soft-threshold partial least squares
SVM-DA	Support vector machines for discriminant analysis
SVMR	Support vector machines regression
SVM-RFE	Support vector machine recursive feature elimination
TN	True negative
TP	True positive
URXY	Univariate regression between each column of matrix **X** with **y**
UV	Ultraviolet spectroscopy
UVE	Uninformative variable elimination
VIP	Variable importance on projection
Vis/NIR	Visible and near-infrared spectroscopy
WGHT	Weights of NIPALS
XRD	X-ray diffraction
XRF	X-ray fluorescence spectrometry

LIST OF FIGURES

CHAPTER I

Figure I.1. Tree map chart of variable selection publications in Web of Science categories referent to a total of 13510 publications (accessed July 10, 2019)..... 33

Figure I.2. Bar plot with the publications about variable selection over the years of 1995 and 2019 in Web of Science database ((accessed July 10, 2019). 33

CHAPTER II

Figure II.1 General scheme of variable selection steps using OPS algorithm. (1) Informative vectors are obtained; (2) the data matrix is differentiated in according to the corresponding absolute values of the informative vector elements; (3) the differentiated variables are sorted in descending order; (4) an initial subset of variables (window) is defined to build and evaluate the first model; (5) this initial subset is extended by the addition of a fix number of variables (increment) over the window, and further increments are added until all, or some percentage of variables are taken into account; (6) cross-validation parameters are calculated for each model; (7) the variable subsets are compared using the quality parameters calculated during validations, and the best variables subset is defined. The different colors represent the weight of the variables and were randomly assigned. 43

Figure II.2 Summary of all informative vector options to be used in new OPS algorithms for variable selection. REG: regression coefficients; COR: correlation between each column of matrix **X** with **y**; SQR: residual information of the reconstructed matrix with *h* latent variables; NAS: net analyte signal; VIP: variable importance on projection; URXY: univariate regression between each column of matrix **X** with **y**; WGHT: weights; COV: covariance procedures; PRODALL: the product of all single vectors simultaneously. 45

Figure II.3 Scheme of the *feed*OPS algorithm where the pre-selected variables return to a new selection run until specific criteria are achieved. *RMSECV*: root mean square error of cross-validation, *n*: loop counter. 47

Figure II.4 Scheme of *i*OPS algorithm where the variable selection is performed in each interval; subsequently, these selected variables are ordered, and a new selection is performed to find the best set of variables. (1) The array is subdivided; (2) *auto*OPS or *feed*OPS is applied in each interval; and a new matrix is created with the selected variables in each interval (3) *auto*OPS or *feed*OPS is applied in the new matrix; (4) the best set of variables is reached. ... 48

Figure II.5 Scheme of the new OPS algorithms (*auto*OPS, *feed*OPS, and *i*OPS) combined with the four vector options. Names of OPS options derived by the combination of new OPS approaches with all vector options are arranged in the lines for the arrows. 49
Figure II.6 (A1) Minimum *sc* values obtained for all informative vectors used in the *auto*OPS algorithm for the XRF dataset. The selected vector was detached as dashed red line. (A2) Detailed results of variable subsets obtained by the *auto*OPS algorithm using the SQR × URXY vector and your respective *RMSECV* and R_{CV} values for the XRF dataset. The dashed red line means the selected subset. Black spheres are *RMSECV* values showed in the left axis, and black squares are R_{CV} values showed in the right axis. (B) *Feed*OPS typical plot (number of selection runs and *RMSECV* values) for the voltammetry dataset. The dashed red line indicates the loop where the convergence was attained, and the set of selected variables was obtained. (C1) *i*OPS plot for the GC dataset with *RMSECV* values obtained by PLS models built using the selected variables in each interval of the full array (first variable selection step). (C2) *i*OPS plot for the GC dataset with *RMSECV* values obtained by PLS models built using the subsets of the merged matrix built from selected variables in each interval. A dashed red line shows the subset where the best set of variables were selected. A solid red line represents the *RMSECV* of full model in each subplot of *i*OPS. *sc*: selection criterion; *RMSECV*: root mean square error of cross-validation. 57
Figure II.7 Number of selected variables placed in the left-side axis (light gray bar) and *RMSEP* values placed in the right-side axis (black spheres). The informative vector and the *h*Mod are placed in bars and black spheres, respectively. (A) NIR – boiling point at 50% recovery, (B) NIR – cetane number, (C) NIR – density, (D) NIR – freezing temperature of the fuel, (E) NIR – total aromatics, (F) NIR – viscosity, (G) Fluorescence – catechol, (H) Fluorescence – hydroquinone, (I) Fluorescence – indole, (J) Fluorescence – resorcinol, (K) Fluorescence – *L*-tryptophane, (L) Fluorescence – *DL*-tyrosine, (M) Simulated – analyte 01, (N) Simulated – analyte 02, (O) Simulated 03 – analyte 03, (P) Simulated – analyte 04, (Q) Raman – active substance in *Escitalopram* ® tablets, (R) GC – flash point, (S) GC – freeze point, (T) GC – specific gravity, and (U) QSAR – *in vitro* inhibition activity. *RMSEP*: root mean square error of prediction, *h*Mod: number of latent variables of the model. *NC: not calculated. 59
Figure II.8 *RMSEP* values of variable selection models (light gray bar) and the number of selected variables (dark gray bar). The values outside the bar are the percent of decrease (negative values) or increase (positive values) of *RMSEP* values or the number of variables regarding full model represented by the horizontal line on zero. (A) Voltammetry – ascorbic acid (μmol L^{-1}), (B) Fluorescence – catechol (μmol L^{-1}), (C) MS – ethanol (vol - %), (D) Raman

– iodine value (g I_2 100g fat^{-1}), (E) NIR – lignin (% w/w), (F) UV – phorbol esters (mg g^{-1}), (G) NMR – pentanol (%), (H) GC – overall quality, (I) Vis/NIR – xylene (weight percent - %), (J) XRF – Ni (%), (K) QSAR – MMP-2 *p*IC_{50}, and (L) MS – peroxide value (meq Kg^{-1}). *RMSEP*: root mean square error of prediction. 64

Figure II.9 (A) Voltammetry – ascorbic acid, (B) Fluorescence – catechol, (C) MS – ethanol, (D) Raman – iodine value, (E) NIR – lignin, (F) UV – phorbol esters, (G) NMR – pentanol, (H) GC – overall quality, (I) Vis/NIR – xylene, (J) XRF – Ni, (K) Autoscale QSAR – MMP-2 pIC50, and (L) MS – peroxide value with the variables selected (vertical lines) by OPS, GA, *i*SPA, and rPLS. The top graph in each subplot is the frequency that a variable was selected by the four variable selection methods. *Variables selected by the model with lowest *RMSEP*.. 65

CHAPTER III

Figure III.1. Algorithm flowchart of the core of OPS method. 73

Figure III.2. Algorithm flowchart of (A) *auto*OPS, (B) *feed*OPS, and (C) *i*OPS methods. ... 73

Figure III.3. Algorithm flowchart of (A) GA, (B) *i*SPA, and (C) rPLS methods. 75

Figure III.4. Comparison of running times of *auto*OPS, *feed*OPS, *i*OPS, GA, *i*SPA, and rPLS algorithms in minutes (A) and in normalized time by Euclidean norm (B). 78

Figure III. 5. Pareto chart for the standardized main effects in the central composite design of *auto*OPS (A), *feed*OPS (B), *i*OPS (C), GA (D), *i*SPA (E), and rPLS (F). Red line is the significance level (α) at 0.05. Linear (L) and quadratic (Q) coefficients. Number of latent variables (*nlvs*)..... 80

CHAPTER IV

Figure IV.1. Scheme of **X** matrix (*i* samples and *j* variables), Class vector, and **Y** binary matrix (*i* samples and *n* classes) in PLS-DA. 89

Figure IV.2. Parameters of decision making in classification models. *TP*: true positive; *FP*: false positive; *FN*: false negative; *TN*: true negative. (●) samples of class 1 and (■) samples of other class. 90

Figure IV.3. Scheme of (A) multiclass and (B) one-vs-all strategies to obtaining selected variables for classification models using OPSDA. **X** matrix (*i* samples and *j* variables), **Y** binary matrix (*i* samples and *n* classes), and varsel (the set of selected variables). 92

Figure IV.4. Number of selection runs in *feedOPSDA* in NIR dataset. The red arrow indicates the loop where the convergence was attained, and a set of selected variables was obtained... 95

Figure IV.5. Cross-validation **Y** prediction values for NIR dataset in full-PLS-DA (A) and OPSDA*pls* (D). Classification by full-PLS-DA (B), OPSDA*pls* (C), full-SVM-DA (E), and

OPSDA*svm* (F) models. Filled red circles (●) and empty red circles (○) are samples of class A. Filled black squares (■) and empty black squares (□) are samples of class B. Filled shapes refer to cross-validation samples, and the empty ones refer to test samples. In (A) and (D) the horizontal dashed black line indicates the threshold selected in PLS-DA method. In (B), (C), (E) and (F), indicates simply a class mark. ErrorCV: classification error in cross-validation. 97

Figure IV.6. (A) NIR spectra and (B) XRD diffractograms with global selected variables by using the *feed*OPSDA with the multiclass strategy. 97

Figure IV.7. Number of selection runs in *feedOPSDA* for each class in XRD dataset. The red arrow indicates the loop where the convergence was attained, and the set of selected variables was obtained. 98

Figure IV.8. Cross-validation **Y** prediction values for XRD dataset in full-PLS-DA (A-C) and OPSDA*pls* (F-H). In each PLS-DA cross-validation plot, the red circles (●) are samples that refer to the class showed in the y-axis. Black squares (■) are samples that refer to other classes. In (A-C) and (F-H) the horizontal dashed black line indicates the threshold selected in PLS-DA method. Classification by full-PLS-DA (D), full-SVM-DA (E), OPSDA*pls* (I), and OPSDA*svm* (J) models. Black diamonds (♦ and ◊) represents class S, red diamonds (♦ and ◊) represents class I, and blue diamonds (♦ and ◊) represents class R. In (D-E) and (I-J), filled shapes refer to training samples and the empty ones refer to test samples; the horizontal dashed black line merely indicates a class mark. 100

Figure IV.9. Number of selection runs in *feedOPSDA* for each class in XRD dataset. The red arrow indicates the loop where the convergence was attained, and the set of selected variables was obtained. 101

Figure IV.10. Cross-validation **Y** prediction values for MS dataset in full-PLS-DA (A-D) and OPSDA*pls* (E-H). In each PLS-DA cross-validation plot, the red circles (●) are samples that refer to the class showed in the y-axis. Black squares (■) are samples that refer to other classes. In (A-H) the horizontal dashed black line indicates the threshold selected in PLS-DA method. Classification by full-PLS-DA (I), full-SVM-DA (K), OPSDA*pls* (J), and OPSDA*svm* (L) models. Black diamonds (♦ and ◊) represents class A, red diamonds (♦ and ◊) represents class B, blue diamonds (♦ and ◊) represents class C, and green diamonds (♦ and ◊) represents class D. In (I-L), filled shapes refer to training samples and the empty ones refer to test samples; the horizontal dashed black line merely indicates a class mark. 103

Figure IV.11. MS spectra with selected variables by OPSDA for class A (A), class B (B), class C (C), and class D (D). 104

LIST OF TABLES

CHAPTER I

Table I.1. Variable selection methods and their classification according to how they combine the selection algorithm and the model building. 32

CHAPTER II

Table II.1 Information about datasets used to perform validation of new strategies of OPS algorithm. 51

Table II.2 Best OPS model, informative vector and the number of latent variables used to perform the selection (*h*OPS) obtained for each dataset. 56

Table II.3 Performance parameters of PLS models obtained using all variables and those selected by OPS, GA, *i*SPA, and rPLS algorithms. 60

CHAPTER III

Table III.1. Factors with coded and decoded levels for five-level CCD designs. 76

Table III.2. Five-level CCD and time results in minutes for variable selection methods. 79

Table III.3. Profile report for OPS, GA, *i*SPA and rPLS with percent of time spent of cross-validation and PLS algorithms. 81

CHAPTER IV

Table IV.1. Statistical parameters of PLS-DA and SVM-DA models using all variables and selected variables by *feed*OPSDA for NIR dataset. 96

Table IV.2. Statistical parameters of PLS-DA and SVM-DA models using all variables and selected variables by *feed*OPSDA for XRD dataset. 99

Table IV.3. Statistical parameters of PLS-DA and SVM-DA models using all variables and selected variables by *feed*OPSDA for MS dataset. 102

LIST OF EQUATIONS

CHAPTER I

$\mathbf{w}_i = \mathbf{X}^t\mathbf{y} / \mathbf{y}^t\mathbf{y}$ **(I.1)** 28

$\mathbf{w}_i = \mathbf{w}_i / \|\mathbf{w}_i\|$ **(I.2)** 28

$\mathbf{t}_i = \mathbf{X}\mathbf{w}_i$ **(I.3)** 28

$\mathbf{p}_i = \mathbf{X}^t\mathbf{t}_i / \mathbf{t}_i^t\mathbf{t}_i$ **(I.4)** 28

$q_i = \mathbf{y}^t\mathbf{t}_i / \mathbf{t}_i^t\mathbf{t}_i$ **(I.5)** 28

$\mathbf{X} = \mathbf{X} - \mathbf{t}_i\mathbf{p}_i^t$ **(I.6)** 29

$\mathbf{y} = \mathbf{y} - \mathbf{t}_i q_i^t$ **(I.7)** 29

$\mathbf{b} = \mathbf{W}^t(\mathbf{P}\mathbf{W}^t)^{-1}q$ **(I.8)** 29

$\mathbf{s} = \mathbf{X}^t\mathbf{y}$ **(I.9)** 29

$\mathbf{r}_i = \mathbf{s}$ **(I.10)** 29

$\mathbf{t}_i = \mathbf{X}\mathbf{r}_i$ **(I.11)** 29

$\mathbf{t}_i = \mathbf{t}_i / \|\mathbf{t}_i\|$ **(I.12)** 29

$\mathbf{r}_i = \mathbf{r}_i / \|\mathbf{r}_i\|$ **(I.13)** 29

$\mathbf{p}_i = \mathbf{X}^t\mathbf{t}_i$ **(I.14)** 29

$q_i = \mathbf{y}^t\mathbf{t}_i$ **(I.15)** 29

$\mathbf{s} = \mathbf{s} - \mathbf{P}(\mathbf{P}^t\mathbf{P})^{-1}\mathbf{s}$ **(I.16)** 29

$\mathbf{b} = \mathbf{R}q$ **(I.17)** 29

$\mathbf{X} = \mathbf{U}\mathbf{R}\mathbf{V}^t$ **(I.18)** 30

$\mathbf{v}_1 = \mathbf{X}^t\mathbf{y} / \|\mathbf{X}^t\mathbf{y}\|$ **(I.19)** 30

$\mathbf{u}_1 = \mathbf{X}\mathbf{v}_1 / \|\mathbf{X}\mathbf{v}_1\|$ **(I.20)** 30

$\alpha_1 = \|\mathbf{X}\mathbf{v}_1\|$ **(I.21)** 30

$\mathbf{y} = \mathbf{y} - \mathbf{u}_1(\mathbf{u}_1^t\mathbf{y})$ **(I.22)** 30

$\mathbf{v}_i = \mathbf{X}^t\mathbf{y} / \|\mathbf{X}^t\mathbf{y}\|$ **(I.23)** 30

$\mathbf{u}_i = \mathbf{X}\mathbf{v}_i$ **(I.24)** 30

$\gamma_{i-1} = \mathbf{U}^t\mathbf{u}_i$ **(I.25)** 30

$\mathbf{u}_i = \mathbf{u}_i - \mathbf{U}(\mathbf{U}^t\mathbf{u}_i) / \left\|\mathbf{u}_i - \mathbf{U}(\mathbf{U}^t\mathbf{u}_i)\right\|$ **(I.26)** 30

$\alpha_i = \left\|\mathbf{u}_i - \mathbf{U}(\mathbf{U}^t\mathbf{u}_i)\right\|$ **(I.27)** 30

$\mathbf{y} = \mathbf{y} - \mathbf{u}_i(\mathbf{u}_i^t\mathbf{y})$ **(I.28)** 30

$\mathbf{b} = \mathbf{X}^+\mathbf{y}$ **(I.29)** 31

$\mathbf{X}^+ = \mathbf{V}\mathbf{R}^{-1}\mathbf{U}^t$ **(I.30)** 31

$\mathbf{b} = \mathbf{V}\mathbf{R}^{-1}\mathbf{U}^t\mathbf{y}$ **(I.31)** 31

CHAPTER II

$RMSE = \sqrt{\sum_i^{I_m}(y_i - \hat{y}_i)^2} \Big/ I_m$ **(II.1)** 41

$R = \sum_{i=1}^{I_m}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y}) \Big/ \sqrt{\sum_{i=1}^{I_m}(\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_{i=1}^{I_m}(y_i - \bar{y})^2}$ **(II.2)** 41

$\hat{\mathbf{b}}_j = (\mathbf{x}_j^t\mathbf{x}_j)^{-1}\mathbf{x}_j^t\mathbf{y}$
$\hat{\mathbf{y}}_j = \mathbf{x}_j\hat{\mathbf{b}}_j$
$\mathbf{ey}_j = \mathbf{y} - \hat{\mathbf{y}}_j$ **(II.3)** 44
$\mathbf{urxy}_j = \hat{b}_j(2) / (\mathbf{ey}_j^t\mathbf{ey}_j)$

$sc_i = RMSECV_i / R_{CV\,i}$ **(II.4)** 45

$rd = RMSECV_n - RMSECV_{n-1} / RMSECV_n$ **(II.5)** 47

CHAPTER III

$\mathbf{b} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$ **(III.1)** 76

CHAPTER IV

$Sensitivity = \frac{TP}{(TP + FN)}$ **(IV.1)** 89

$Specificity = \frac{TN}{(TN + FP)}$ **(IV.2)** 89

$Error = \frac{FP + FN}{(TP + TN + FP + FN)}$ **(IV.3)** 90

$rd = \frac{ErrorCV_n - ErrorCV_{n-1}}{ErrorCV_n}$ **(IV.4)** 91

SUMMARY

GUIDELINES **22**

GENERAL INTRODUCTION **23**

References **24**

CHAPTER I **26**

Partial Least Squares **27**

Notation *28*

NIPALS *28*

SIMPLS *29*

BIDIAGONALIZATION *30*

Variable Selection **31**

References **34**

CHAPTER II **38**

Abstract **39**

Introduction **40**

Theory Background **41**

Model evaluation *41*

OPS theory *42*

Automatic OPS (autoOPS) *43*

Feedback OPS (feedOPS) *46*

OPS Intervals (iOPS) *48*

Experimental **49**

Modeling *49*

Datasets *53*

Results and Discussion **55**

Algorithms *55*

Modeling *58*

Conclusion **66**

References **66**

CHAPTER III **69**

Abstract **70**

Introduction **71**

Variable Selection Methods **72**

Ordered Predictors Selection (OPS) *72*

Genetic Algorithm (GA) *72*

Intervals Successive Projections Algorithm (iSPA) *74*

Recursive Weighted Partial Least Squares (rPLS) *74*

Experimental **74**

Simulated Datasets *75*

Algorithms Parameters *76*
Results and Discussion **77**
Conclusion **82**
References **83**

CHAPTER IV **85**
Abstract **86**
Introduction **87**
Theory background **88**
Partial Least Squares for Discriminant Analysis (PLS-DA) *88*
Model Evaluation *89*
Ordered Predictors Selection for Discriminant Analysis (OPSDA) *90*
Experimental **92**
Near-Infrared Spectroscopy (NIR) *93*
X-Ray Diffraction (XRD) *94*
Mass Spectrometry (MS) *94*
Results and Discussion **95**
Near-Infrared Spectroscopy (NIR) *95*
X-Ray Diffraction (XRD) *98*
Mass Spectrometry (MS) *100*
Conclusion **104**
References **104**

CONCLUDING REMARKS **109**

GUIDELINES

This thesis is about an upgrade of the ordered predictors selection (OPS), which is a method of variable selection, and it is divided into four chapters. The outline of this work is as follows. Chapter I is a brief description of partial least squares (PLS) algorithms and an overview of variable selection. Chapter II describes three new OPS methods to apply variable selection in multivariate regression: automatic OPS (*auto*OPS), feedback OPS (*feed*OPS), and interval OPS (*i*OPS). Chapter III presents a study about the computational running time of variable selection methods. In Chapter IV, new OPS methods to select variables in discriminant analysis (OPSDA) is presented, covering three approaches: automatic OPS for discriminant analysis (*auto*OPSDA); feedback OPS for discriminant analysis (*feed*OPSDA), and interval OPS for discriminant analysis (*i*OPSDA).

GENERAL INTRODUCTION

Multivariate statistical methods have been increasingly used in chemistry in the last 25 years, due to the greater availability of analytical instruments, software, and computers [1]. Multivariate regression and pattern recognition are the major areas of chemometrics related to analytical chemistry and they are widely used in chemical problem solving [2]. The first one describes the quantitative relationship between the dependent and independent variables [3,4]. The second depends on what is known a priori. When no information about samples grouping is required, the methods are called unsupervised pattern recognition. On the other hand, when the group assignments are known (classes), the methods are named supervised pattern recognition or classification methods. There are several multivariate regression and supervised pattern recognition methods available in the literature. However, two methods were used in this work, being then partial least squares regression (PLS) and its variant partial least squares for discriminant analysis (PLS-DA).

PLS and PLS-DA can deal with a large number of highly correlated independent variables, band overlaps, and experimental noise [5,6]. However, in some situations, the models cannot provide a satisfactory prediction or classification. For this reason, there are several approaches to search for the most significant variables in both regression and classification problems. Variable selection has become a fundamental tool in many different research areas. It reduces data dimensionality, minimizing the influence of irrelevant variables, improving the model´s quality, and simplifying its interpretation [3,7]. Several feature-selection methods [8–13] have been applied in several works in the literature [14–21]. Nevertheless, most selection methods are not generalized or efficient for all types of datasets. Different analytical techniques provide distinctive predictor types, and each technique has some well-defined peculiarities [22].

Teófilo et al. presented in 2009 the first version of the ordered predictors selection (OPS) method [22], with the advantage of being simple and efficient for the selection of any variable type, and it has its potential recognized in the literature [23–25]. However, the original OPS method was projected to select variables to build only regression models. It is limited to just one variable selection run, which turns out to be limitations of the original version when searching for the best set of variables. Therefore, the main goal of this work was to make a more versatile OPS, developing new OPS methods for regression and discriminant analysis.

The specific aims of this work are:

- To develop three new OPS approaches for multivariate regression;

- To compare the new OPS approaches with the first version of OPS algorithm (OPS*v*1);
- To apply the new OPS approaches in several types of dataset (sparse or non-sparse);
- To compare the new OPS approaches with other variable selection methods;
- To evaluate the computational running time of the new OPS approaches;
- To develop new OPS methods for discriminant analysis (OPSDA);
- To apply the OPSDA in several datasets with different numbers of classes;
- To build classification models using different methods with the selected variables by OPSDA.

References

[1] M.M.C. Ferreira, Quimiometria – Conceitos, Métodos e Aplicações., Unicamp, Editora Unicamp, Campinas, SP, SP, 2015.
[2] P.K. Hopke, The evolution of chemometrics, Anal. Chim. Acta. 500 (2003) 365–377. doi:10.1016/S0003-2670(03)00944-9.
[3] C.M. Andersen, R. Bro, Variable selection in regression-a tutorial, J. Chemom. 24 (2010) 728–737. doi:10.1002/cem.1360.
[4] B. Nadler, R.R. Coifman, The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration, J. Chemom. 19 (2005) 107–118. doi:10.1002/cem.915.
[5] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130. doi:10.1016/S0169-7439(01)00155-1.
[6] S. Wold, J. Trygg, A. Berglund, H. Antti, Some recent developments in PLS modeling, Chemom. Intell. Lab. Syst. 58 (2001) 131–150. doi:10.1016/S0169-7439(01)00156-3.
[7] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta. 667 (2010) 14–32. doi:10.1016/j.aca.2010.03.048.
[8] R. Leardi, A. Lupiáñez González, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemom. Intell. Lab. Syst. 41 (1998) 195–207. doi:10.1016/S0169-7439(98)00051-3.
[9] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Elimination of uninformative variables for multivariate calibration., Anal. Chem. 68 (1996) 3851–3858. doi:10.1021/ac960321m.
[10] Å. Rinnan, M. Andersson, C. Ridder, S.B. Engelsen, Recursive weighted partial least squares (rPLS): an efficient variable selection method using PLS, J. Chemom. 28 (2014) 439–447. doi:10.1002/cem.2582.
[11] R. Leardi, L. Nørgaard, Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, J. Chemom. 18 (2004) 486–497. doi:10.1002/cem.893.
[12] A.A. Gomes, R.K.H. Galvão, M.C.U. Araújo, G. Véras, E.C. da Silva, The successive projections algorithm for interval selection in PLS, Microchem. J. 110 (2013) 202–208. doi:10.1016/j.microc.2013.03.015.
[13] L. Norgaard, 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.

doi:10.1366/0003702001949500.
[14] C.V. Di Anibal, M.P. Callao, I. Ruisánchez, 1H NMR variable selection approaches for classification. A case study: The determination of adulterated foodstuffs, Talanta. 86 (2011) 316–323. doi:https://doi.org/10.1016/j.talanta.2011.09.019.
[15] W. Fan, H. Li, Y. Shan, H. Lv, H. Zhang, Y. Liang, Classification of vinegar samples based on near infrared spectroscopy combined with wavelength selection, Anal. Methods. 3 (2011) 1872–1876. doi:10.1039/c1ay05101f.
[16] J.P.M. Andries, Y. Vander Heyden, L.M.C. Buydens, Predictive-Property-Ranked Variable Reduction with Final Complexity Adapted Models in Partial Least Squares Modeling for Multiple Responses, Anal. Chem. 85 (2013) 5444–5453. doi:10.1021/ac400339e.
[17] A.M.K. Pedro, M.M.C. Ferreira, Nondestructive Determination of Solids and Carotenoids in Tomato Products by Near-Infrared Spectroscopy and Multivariate Calibration, Anal. Chem. 77 (2005) 2505–2511. doi:10.1021/ac048651r.
[18] R.M. Balabin, S. V. Smirnov, Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data, Anal. Chim. Acta. 692 (2011) 63–72. doi:10.1016/j.aca.2011.03.006.
[19] I.R.N. de Oliveira, J. V. Roque, M.P. Maia, P.C. Stringheta, R.F. Teófilo, New strategy for determination of anthocyanins, polyphenols and antioxidant capacity of Brassica oleracea liquid extract using infrared spectroscopies and multivariate regression, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 194 (2018) 172–180. doi:10.1016/j.saa.2018.01.006.
[20] C. Assis, R.S. Ramos, L.A. Silva, V. Kist, M.H.P. Barbosa, R.F. Teófilo, Prediction of Lignin Content in Different Parts of Sugarcane Using Near-Infrared Spectroscopy (NIR), Ordered Predictors Selection (OPS), and Partial Least Squares (PLS), Appl. Spectrosc. 71 (2017) 2001–2012. doi:10.1177/0003702817704147.
[21] A. Baldovin, W. Wu, V. Centner, D. Jouan-Rimbaud, D.L. Massart, L. Favretto, A. Turello, Feature selection for the discrimination between pollution types with partial least squares modelling, Analyst. 121 (1996) 1603–1608. doi:10.1039/AN9962101603.
[22] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression, J. Chemom. 23 (2009) 32–48. doi:10.1002/cem.1192.
[23] M. Goodarzi, Y. Vander Heyden, S. Funar-Timofei, Towards better understanding of feature-selection or reduction techniques for Quantitative Structure–Activity Relationship models, TrAC Trends Anal. Chem. 42 (2013) 49–63. doi:10.1016/j.trac.2012.09.008.
[24] J. Yuan, S. Yu, T. Zhang, X. Yuan, Y. Cao, X. Yu, X. Yang, W. Yao, QSPR models for predicting generator-column-derived octanol/water and octanol/air partition coefficients of polychlorinated biphenyls, Ecotoxicol. Environ. Saf. 128 (2016) 171–180. doi:10.1016/j.ecoenv.2016.02.022.
[25] C.S.W. Miaw, C. Assis, A.R.C.S. Silva, M.L. Cunha, M.M. Sena, S.V.C. de Souza, Determination of main fruits in adulterated nectars by ATR-FTIR spectroscopy combined with multivariate calibration and variable selection methods, Food Chem. 254 (2018) 272–280. doi:10.1016/j.foodchem.2018.02.015.

Chapter I

Literature review

Partial Least Squares

In chemometrics, the partial least squares (PLS) is a consolidated statistical tool for modeling the linear relationship between multivariate measurements [1,2]. PLS is a method to predict a set of dependent variables (**Y**) from a set of independent variables or predictors (**X**) [3]. PLS regression finds a linear model by projecting **X** and **Y** data into a new space. This method will try to find the multidimensional direction in the **X** space that explains the maximum multidimensional variance direction in the **Y** space [4], *i.e.*, a latent variable approach to modeling the covariance structures in these two spaces [3]. Because of this, the PLS approach is known as bilinear factor model [1,4].

Partial least squares for discriminant analysis (PLS-DA) is a variant used when the **Y** variable is categorical [5].

PLS is particularly suited when the matrix **X** has more variables than observations, *i.e.*, more columns than rows [6,7]. The great advantage of PLS is to analyze data with numerous **X**-variables strongly collinear, noisy, and simultaneously model several response variables [8,9], enabling the application to several types of datasets. PLS has been successfully applied to the quantitative analyses in a series of chemical data such as ultraviolet [10], near-infrared [11–14], mid-infrared [15], nuclear magnetic resonance [16], chromatographic [17–19], quantitative structure-activity / structure-property relationship [20–22], electrochemical [23], X-ray fluorescence [24], and laser-induced breakdown spectroscopy [25].

The pioneering work in PLS was originated by Herman Wold (1975) in econometrics field. Around 1980, the chemometric version of PLS was developed by Svante Wold, Herman Wold's son and Harald Martens to better suit to data from science and technology. This modified method shows to be useful to deal with complicated datasets where ordinary regression was difficult or impossible to apply [1].

The regression problem is one of the most common data-analytical problems in science and technology, *i.e.*, how to model one or several dependent variables, responses, **Y**, employing a set of predictor variables, **X** [1]. PLS regression builds a regression model in a single step, which both information of **X** and **Y** are considered during the data decomposition and compression. In this step, each latent variable is obtained by maximizing the covariance between the **X** and **Y** scores [9]. In this thesis is referred only to PLS1 regression with only one dependent variable in **Y**, *i.e.*, **y**.

A common question is why PLS is a "partial" least squares regression. This term has been referred to as iterative, focusing on the iterative nature of the procedure of estimating weights and latent variables [2,26]. The first use of the term PLS occurred referring to nonlinear iterative partial least squares (NIPALS) estimation procedure [27]. From this, the term "partial" has its roots in the NIPALS procedures, which implemented the idea that the model parameters can be estimated in parts [28].

About PLS calculations, several variations of PLS algorithms were developed over the years. Among the most used NIPALS [2,29], SIMPLS [30] and the bidiagonalization algorithm [31,32], that are described in the following sections.

Notation

Matrices are represented by bold upper-case letters (**A**, **B**, **C**), vectors by bold lower-case letters (**a**, **b**, **c**), and scalars by italic lower-case letters (*a*, *b*, *c*). Superscripts ^{t}, $^{-1}$, and $^{+}$ represent transpose, inverse, and Moore-Penrose generalized inverse operations, respectively. The Euclidean norm of **a** is denoted by $\|\mathbf{a}\|$. The identity matrix is represented as **I**. The independent variables are designed as **X** ($n \times m$), where each row contains the measurement for a given sample, and each column the measurement for a given variable. For each sample, there is a dependent variable *y*, which is represented by a column vector **y** ($n \times 1$).

NIPALS

The NIPALS [2,29] was the first algorithm used in PLS regression, and it can be summarized as follows:

For i = 1, 2, ..., h latent variables, calculate **w** (weights for **X)**, **t** (scores for **X)**, q (loadings for **y)**, and **p** (loadings for **X):**

$$\mathbf{w}_i = \mathbf{X}^t\mathbf{y} / \mathbf{y}^t\mathbf{y} \quad \textbf{(I.1)}$$

$$\mathbf{w}_i = \mathbf{w}_i / \|\mathbf{w}_i\| \quad \textbf{(I.2)}$$

$$\mathbf{t}_i = \mathbf{X}\mathbf{w}_i \quad \textbf{(I.3)}$$

$$\mathbf{p}_i = \mathbf{X}^t\mathbf{t}_i / \mathbf{t}_i^t\mathbf{t}_i \quad \textbf{(I.4)}$$

$$q_i = \mathbf{y}^t\mathbf{t}_i / \mathbf{t}_i^t\mathbf{t}_i \quad \textbf{(I.5)}$$

Deflate **X** and **y** by subtracting the computed latent vectors:

$$\mathbf{X} = \mathbf{X} - \mathbf{t}_i \mathbf{p}_i^t \quad \textbf{(I.6)}$$

$$\mathbf{y} = \mathbf{y} - \mathbf{t}_i q_i^t \quad \textbf{(I.7)}$$

After deflating steps, **w**, **t**, q, and **p** will be calculated for the next latent variable by equations (I.1) to (I.5) until achieve $i = h$.

Store $\mathbf{W} = (\mathbf{w}_1, ..., \mathbf{w}_h), \mathbf{T} = (\mathbf{t}_1, ..., \mathbf{t}_h), \mathbf{P} = (\mathbf{p}_1, ..., \mathbf{p}_h), \mathbf{q} = (q_1, ..., q_h)$, and the regression coefficients can be calculated as:

$$\mathbf{b} = \mathbf{W}^t (\mathbf{P}\mathbf{W}^t)^{-1} q \quad \textbf{(I.8)}$$

SIMPLS

The SIMPLS algorithm was developed by De Jong [30] with the advantage that is not necessary to deflate **X** or **y**. This algorithm is initialized computing **s**:

$$\mathbf{s} = \mathbf{X}^t \mathbf{y} \quad \textbf{(I.9)}$$

For $i = 1, 2, ..., h$ latent variables, calculate **r** (weights for **X)**, **t** (scores for **X)**, q (loadings for **y)**, and **p** (loadings for **X):**

$$\mathbf{r}_i = \mathbf{s} \quad \textbf{(I.10)}$$

$$\mathbf{t}_i = \mathbf{X}\mathbf{r}_i \quad \textbf{(I.11)}$$

$$\mathbf{t}_i = \mathbf{t}_i / \|\mathbf{t}_i\| \quad \textbf{(I.12)}$$

$$\mathbf{r}_i = \mathbf{r}_i / \|\mathbf{r}_i\| \quad \textbf{(I.13)}$$

$$\mathbf{p}_i = \mathbf{X}^t \mathbf{t}_i \quad \textbf{(I.14)}$$

$$q_i = \mathbf{y}^t \mathbf{t}_i \quad \textbf{(I.15)}$$

Store $\mathbf{R} = (\mathbf{r}_1, ..., \mathbf{r}_h), \mathbf{T} = (\mathbf{t}_1, ..., \mathbf{t}_h), \mathbf{P} = (\mathbf{p}_1, ..., \mathbf{p}_h), \mathbf{q} = (q_1, ..., q_h)$ and project **s** on a subspace orthogonal to **P**:

$$\mathbf{s} = \mathbf{s} - \mathbf{P}(\mathbf{P}^t \mathbf{P})^{-1} \mathbf{s} \quad \textbf{(I.16)}$$

After the projection step, **r**, **t**, q, and **p** will be calculated for the next latent variable by equations (I.10) to (I.15) until achieve $i = h$. The last step calculates the regression vector:

$$\mathbf{b} = \mathbf{R}q \quad \textbf{(I.17)}$$

BIDIAGONALIZATION

The bidiagonalization algorithm [31,32] consists of writing any **X** matrix as:

$$\mathbf{X} = \mathbf{URV}^t \quad \textbf{(I.18)}$$

where **U** and **V** are matrices with orthonormal columns, and **R** is a bidiagonal matrix. In the bidiagonal algorithm the **y** is incorporated in decomposition of **X**.

Initialize the bidiagonalization algorithm for the first latent variable:

$$\mathbf{v}_1 = \mathbf{X}^t\mathbf{y} / \left\|\mathbf{X}^t\mathbf{y}\right\| \quad \textbf{(I.19)}$$

$$\mathbf{u}_1 = \mathbf{X}\mathbf{v}_1 / \left\|\mathbf{X}\mathbf{v}_1\right\| \quad \textbf{(I.20)}$$

$$\alpha_1 = \left\|\mathbf{X}\mathbf{v}_1\right\| \quad \textbf{(I.21)}$$

Store $\mathbf{V} = \mathbf{v}_1$ and $\mathbf{U} = \mathbf{u}_1$.

$$\mathbf{y} = \mathbf{y} - \mathbf{u}_1(\mathbf{u}_1^t\mathbf{y}) \quad \textbf{(I.22)}$$

After the deflating **y** in equation (I.22), **u** and **v** vectors, α and γ left and right diagonal elements, respectively, can be calculated for $i = 2, ..., h$ latent variables:

$$\mathbf{v}_i = \mathbf{X}^t\mathbf{y} / \left\|\mathbf{X}^t\mathbf{y}\right\| \quad \textbf{(I.23)}$$

$$\mathbf{u}_i = \mathbf{X}\mathbf{v}_i \quad \textbf{(I.24)}$$

$$\gamma_{i-1} = \mathbf{U}^t\mathbf{u}_i \quad \textbf{(I.25)}$$

$$\mathbf{u}_i = \mathbf{u}_i - \mathbf{U}(\mathbf{U}^t\mathbf{u}_i) / \left\|\mathbf{u}_i - \mathbf{U}(\mathbf{U}^t\mathbf{u}_i)\right\| \quad \textbf{(I.26)}$$

$$\alpha_i = \left\|\mathbf{u}_i - \mathbf{U}(\mathbf{U}^t\mathbf{u}_i)\right\| \quad \textbf{(I.27)}$$

$$\mathbf{y} = \mathbf{y} - \mathbf{u}_i(\mathbf{u}_i^t\mathbf{y}) \quad \textbf{(I.28)}$$

Store $\mathbf{V} = (\mathbf{v}_1,...,\mathbf{v}_h)$, $\mathbf{U} = (\mathbf{u}_1,...,\mathbf{u}_h)$ and $\mathbf{R} = \begin{pmatrix} \alpha_1 & \gamma_1 & & \\ & \ddots & \ddots & \\ & & \alpha_{h-1} & \gamma_{h-1} \\ & & & \alpha_h \end{pmatrix}$.

After the deflating step (I.28), **u**, **v**, α, and γ will be calculated for the next latent variable by equations (I.23) to (I.27) until achieve $i = h$.

In bidiagonalization decomposition, **UR** is the scores matrix and $\mathbf{V}^t$ is the loading matrix. A compact formulation of the solution of **y** = **Xb** may be given as:

$$\mathbf{b} = \mathbf{X}^{+}\mathbf{y} \qquad \textbf{(I.29)}$$

Because $\mathbf{U}^t\mathbf{U} = \mathbf{V}^t\mathbf{V} = \mathbf{I}$, the **R** can be written as $\mathbf{R} = \mathbf{U}^t\mathbf{X}\mathbf{V}$, and the Moore-Penrose generalized inverse can be obtained as:

$$\mathbf{X}^{+} = \mathbf{V}\mathbf{R}^{-1}\mathbf{U}^t \qquad \textbf{(I.30)}$$

With the orthogonal decomposition in equation (I.18) one may thus obtain the regression coefficients from (I.29) and (I.30) as:

$$\mathbf{b} = \mathbf{V}\mathbf{R}^{-1}\mathbf{U}^t\mathbf{y} \qquad \textbf{(I.31)}$$

Variable Selection

Multivariate models, such as PLS, have been developed for quantitative or qualitative analysis of data because of their ability to reduce the impact of common problems such as collinearity, band overlaps, and interactions. However, data that contains irrelevant features can negatively influence the regression models. Having irrelevant features in data, like noise variables, it can decrease the accuracy of the models and make the model learn based on irrelevant features. Therefore, removal of the variables in which the noise dominates over the relevant information often leads to better accuracy and performance of the multivariate models. This procedure, called feature or variable selection, has become a necessary method in different domains, because of the increasing dataset dimensions [33–36].

Variable selection is a process to select a subset of relevant features, variables, or predictors, which contribute to improving the model's performance. The main goal is to identify a subset of variables that produce the smallest possible errors when used to perform operations as making quantitative determinations or discriminating between different samples [34,37]. Besides, by selecting a relevant number of features from a dataset, not only a better fitting, a better predictive and simpler model can be built, but also knowledge about the most important variables can help in the interpretation [36]. This approach is considered an important pre-processing procedure in chemometrics, which is widely used to improve the performance of various multivariate methods and algorithms, such as regression and classification methods [33].

There are numerous suggested methods for variable selection in the literature of data analysis and statistics, and it is a challenge to stay updated on all the possibilities. Table I.1

presents some examples of variable selection methods described in the literature based on different criteria and selection mechanisms. Feature selection techniques can be categorized into three classes based on how they combine the selection algorithm and the model building: filter, wrapper, and hybrid or embedded methods [36,37].

Filter techniques are unsupervised methods based only on general features like the correlation with the variable to predict, which reduce variables into a smaller set based on a specified criterion. First, a multivariate model is built, and second, the variable selection is carried out using a simple threshold. Usually, filter methods are fast and easy to compute, but they tend to select redundant variables since they do not take into account the collinearity between variables [33,37].

Table I.1. Variable selection methods and their classification according to how they combine the selection algorithm and the model building.

Method	Reference	Category
Recursive Partial Least Squares (rPLS)	[38]	
Successive Projection Algorithm (SPA)	[39–41]	Filter
Variable Importance in Projection (VIP)	[42]	
Interval Partial Least Squares (*i*PLS)	[43]	
Uninformative Variable Elimination (UVE)	[44]	Wrapper
Genetic Algorithm (GA)	[45–49]	
Ordered Predictors Selection (OPS)	[50,51]	
Interactive Variable Selection (IVS)	[52]	Embedded
Soft-Threshold PLS (ST-PLS)	[53]	

Wrapper methods are based on an optimization criterion to select variables with a double iterative procedure where outer iterations for variable selection and inner iterations of model refitting works iteratively [54]. These methods evaluate subsets of variables which allows detecting the possible interactions between variables. However, they usually spend an intensive computational time and are subject to the risk of overfitting [37].

A method can be called hybrid or embedded when both filter and wrapper methods are used together in selecting a set of relevant variables [36,55]. The variable selection step is an integrated part of a regression algorithm based on one iterative procedure. Then, the embedded methods are typically less time consuming than the wrapper methods [36,37,55].

Variable selection has proven to be a versatile tool for multivariate data analysis and the number of applications coupled with PLS in research fields like statistics, computer science, analytical chemistry, instrumentation is shown in Figure I.1 (source: www.webofknowledge.com with search keywords ("variable selection" or "feature selection") and "partial least squares").

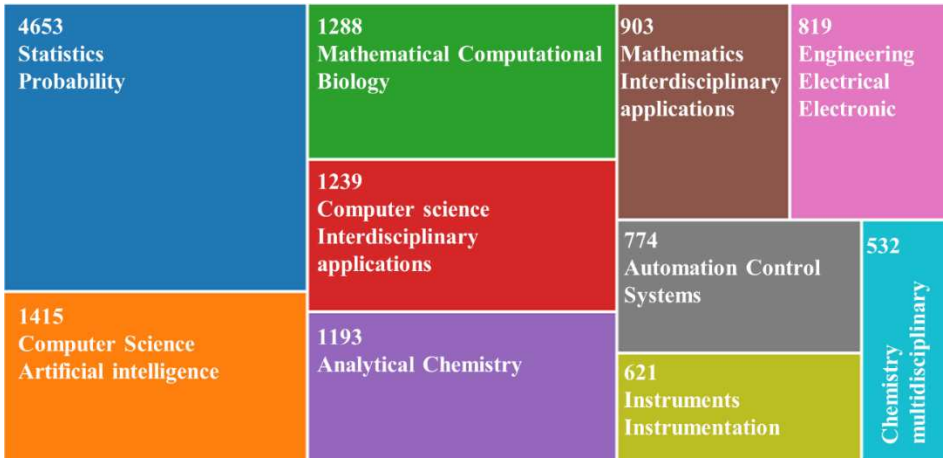

Figure I.1. Tree map chart of variable selection publications in Web of Science categories referent to a total of 13510 publications (accessed July 10, 2019).

Besides that, in Figure I.2 (source: www.webofknowledge.com with search keywords ("variable selection" or "feature selection") and "partial least squares") the number of publications using variable selection with PLS is steadily increasing over the years, showing the importance of this approach.

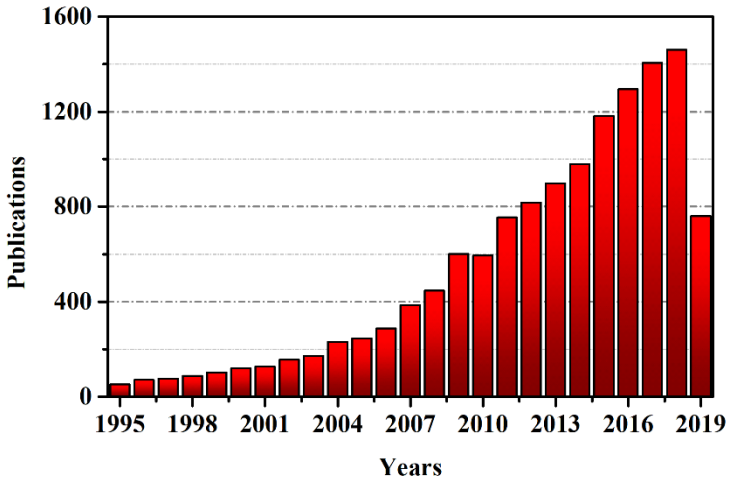

Figure I.2. Bar plot with the publications about variable selection over the years of 1995 and 2019 in Web of Science database ((accessed July 10, 2019).

Although there are several algorithms for feature selection, new methods have been developed and disseminated aiming a better prediction, both in regression and classification. However, most of the variable selection methods are not generalized and efficient for all types of variables. For example, in chemistry, several analytical techniques provide different types of predictors, and each one presents well-defined characteristics and peculiarities [51]. In this context, this work was based on developing a more versatile variable selection method, the ordered predictors selection for regression and classification, expanding its worldwide use and applicability.

References

[1] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130. doi:10.1016/S0169-7439(01)00155-1.
[2] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta. 185 (1986) 1–17. doi:10.1016/0003-2670(86)80028-9.
[3] H. Abdi, Partial least squares regression and projection on latent structure regression (PLS Regression), Wiley Interdiscip. Rev. Comput. Stat. 2 (2010) 97–106. doi:10.1002/wics.51.
[4] A. Höskuldsson, PLS regression methods, J. Chemom. 2 (1988) 211–228. doi:10.1002/cem.1180020306.
[5] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemom. 17 (2003) 166–173. doi:10.1002/cem.785.
[6] P. Gemperline, Pratical Guide to Chemometrics, 2nd ed., Taylor e Francis, New York, 2006. doi:10.1201/9781420018301.
[7] K.R. Beebe, B.R. Kowalski, An Introduction to Multivariate Calibration and Analysis, Anal. Chem. 59 (1987) 1007A-1017A. doi:10.1021/ac00144a725.
[8] S. Wold, J. Trygg, A. Berglund, H. Antti, Some recent developments in PLS modeling, Chemom. Intell. Lab. Syst. 58 (2001) 131–150. doi:10.1016/S0169-7439(01)00156-3.
[9] M.M.C. Ferreira, Quimiometria – Conceitos, Métodos e Aplicações., Unicamp, Editora Unicamp, Campinas, SP, SP, 2015.
[10] J. V Roque, L.A.S. Dias, R.F. Teófilo, Multivariate Calibration to Determine Phorbol Esters in Seeds of Jatropha curcas L. Using Near Infrared and Ultraviolet Spectroscopies, J. Braz. Chem. Soc. 28 (2017) 1506–1516. doi:10.21577/0103-5053.20160332.
[11] G.W.D. Ferreira, J. V. Roque, E.M.B. Soares, I.R. Silva, E.F. Silva, A. A. Vasconcelos, R.F. Teófilo, Temporal decomposition sampling and chemical characterization of eucalyptus harvest residues using NIR spectroscopy and chemometric methods, Talanta. 188 (2018) 168–177. doi:10.1016/j.talanta.2018.05.073.
[12] C. Assis, R.S. Ramos, L.A. Silva, V. Kist, M.H.P. Barbosa, R.F. Teófilo, Prediction of Lignin Content in Different Parts of Sugarcane Using Near-Infrared Spectroscopy (NIR), Ordered Predictors Selection (OPS), and Partial Least Squares (PLS), Appl. Spectrosc. 71 (2017) 2001–2012. doi:10.1177/0003702817704147.
[13] Í.P. Caliari, M.H.P. Barbosa, S.O. Ferreira, R.F. Teófilo, Estimation of cellulose crystallinity of sugarcane biomass using near infrared spectroscopy and multivariate analysis methods, Carbohydr. Polym. 158 (2017) 20–28. doi:10.1016/j.carbpol.2016.12.005.

[14] S. Li, X. Fan, J. Mei, G. Shen, L. Han, Y. Li, X. Liu, Z. Yang, Identification of antibiotic mycelia residues in cottonseed meal using Fourier transform near-infrared microspectroscopic imaging, Food Chem. 293 (2019) 204–212. doi:10.1016/j.foodchem.2019.04.100.

[15] I.R.N. de Oliveira, J. V. Roque, M.P. Maia, P.C. Stringheta, R.F. Teófilo, New strategy for determination of anthocyanins, polyphenols and antioxidant capacity of Brassica oleracea liquid extract using infrared spectroscopies and multivariate regression, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 194 (2018) 172–180. doi:10.1016/j.saa.2018.01.006.

[16] A.P. Vieira, N.A. Portela, Á.C. Neto, V. Lacerda, W. Romão, E.V.R. Castro, P.R. Filgueiras, Determination of physicochemical properties of petroleum using 1H NMR spectroscopy combined with multivariate calibration, Fuel. 253 (2019) 320–326. doi:10.1016/j.fuel.2019.05.028.

[17] J.S. Ribeiro, F. Augusto, T.J.G. Salva, M.M.C. Ferreira, Prediction models for Arabica coffee beverage quality based on aroma analyses and chemometrics, Talanta. 101 (2012) 253–260. doi:10.1016/j.talanta.2012.09.022.

[18] J.S. Ribeiro, F. Augusto, T.J.G. Salva, M.M.C. Ferreira, Prediction models for Arabica coffee beverage quality based on aroma analyses and chemometrics, Talanta. 101 (2012) 253–260. doi:10.1016/j.talanta.2012.09.022.

[19] H. Yu, T. Xie, J. Xie, L. Ai, H. Tian, Characterization of key aroma compounds in Chinese rice wine using gas chromatography-mass spectrometry and gas chromatography-olfactometry, Food Chem. 293 (2019) 8–14. doi:10.1016/j.foodchem.2019.03.071.

[20] E.B. De Melo, Modeling physical and toxicity endpoints of alkyl (1-phenylsulfonyl) cycloalkane-carboxylates using the Ordered Predictors Selection (OPS) for variable selection and descriptors derived with SMILES, Chemom. Intell. Lab. Syst. 118 (2012) 79–87. doi:10.1016/j.chemolab.2012.08.006.

[21] J. Yuan, S. Yu, T. Zhang, X. Yuan, Y. Cao, X. Yu, X. Yang, W. Yao, QSPR models for predicting generator-column-derived octanol/water and octanol/air partition coefficients of polychlorinated biphenyls, Ecotoxicol. Environ. Saf. 128 (2016) 171–180. doi:10.1016/j.ecoenv.2016.02.022.

[22] R. Kiralj, M.M.C. Ferreira, Basic validation procedures for regression models in QSAR and QSPR studies: theory and application, J. Braz. Chem. Soc. 20 (2009) 770–787. doi:10.1590/S0103-50532009000400021.

[23] J. Meng, D.-I. Stroe, M. Ricco, G. Luo, R. Teodorescu, A Simplified Model-Based State-of-Charge Estimation Approach for Lithium-Ion Battery With Dynamic Linear Model, IEEE Trans. Ind. Electron. 66 (2019) 7717–7727. doi:10.1109/TIE.2018.2880668.

[24] S. Nawar, N. Delbecque, Y. Declercq, P. De Smedt, P. Finke, A. Verdoodt, M. Van Meirvenne, A.M. Mouazen, Can spectral analyses improve measurement of key soil fertility parameters with X-ray fluorescence spectrometry?, Geoderma. 350 (2019) 29–39. doi:10.1016/j.geoderma.2019.05.002.

[25] R. Yuan, Y. Tang, Z. Zhu, Z. Hao, J. Li, H. Yu, Y. Yu, L. Guo, X. Zeng, Y. Lu, Accuracy improvement of quantitative analysis for major elements in laser-induced breakdown spectroscopy using single-sample calibration, Anal. Chim. Acta. 1064 (2019) 11–16. doi:10.1016/j.aca.2019.02.056.

[26] H. Martens, T. Naes, T. Naes, Multivariate calibration, John Wiley & Sons, 1992.

[27] R. Rosipal, Nonlinear Partial Least Squares An Overview, in: Chemoinformatics Adv. Mach. Learn. Perspect., IGI Global, 2011: pp. 169–189. doi:10.4018/978-1-61520-911-

8.ch009.
[28] G. Sanchez, PLS path modeling with R, Berkeley Trowchez Ed. 383 (2013) 2013.
[29] D.M. Haaland, E. V. Thomas, Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information, Anal. Chem. 60 (1988) 1193–1202. doi:10.1021/ac00162a020.
[30] S. De Jong, SIMPLS: An alternative approach to partial least squares regression, Chemom. Intell. Lab. Syst. 18 (1993) 251–263. doi:10.1016/0169-7439(93)85002-X.
[31] G.H. Golub, W. Kahan, Calculating the singular values and pseudo-inverse of a matrix, SIAM J. Numer. Anal. 2 (1965) 205–224.
[32] W. Wu, R. Manne, Fast regression methods in a Lanczos (or PLS-1) basis. Theory and applications, Chemom. Intell. Lab. Syst. 51 (2000) 145–161. doi:10.1016/S0169-7439(00)00063-0.
[33] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta. 667 (2010) 14–32. doi:10.1016/j.aca.2010.03.048.
[34] R.M. Balabin, S. V. Smirnov, Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data, Anal. Chim. Acta. 692 (2011) 63–72. doi:10.1016/j.aca.2011.03.006.
[35] A. Höskuldsson, Variable and subset selection in PLS regression, Chemom. Intell. Lab. Syst. 55 (2001) 23–38. doi:10.1016/S0169-7439(00)00113-1.
[36] M. Goodarzi, Y. Vander Heyden, S. Funar-Timofei, Towards better understanding of feature-selection or reduction techniques for Quantitative Structure–Activity Relationship models, TrAC - Trends Anal. Chem. 42 (2013) 49–63. doi:10.1016/j.trac.2012.09.008.
[37] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in Partial Least Squares Regression, Chemom. Intell. Lab. Syst. 118 (2012) 62–69. doi:10.1016/j.chemolab.2012.07.010.
[38] Å. Rinnan, M. Andersson, C. Ridder, S.B. Engelsen, Recursive weighted partial least squares (rPLS): an efficient variable selection method using PLS, J. Chemom. 28 (2014) 439–447. doi:10.1002/cem.2582.
[39] S.F.C. Soares, A.A. Gomes, M.C.U. Araujo, A.R.G. Filho, R.K.H. Galvão, The successive projections algorithm, TrAC Trends Anal. Chem. 42 (2013) 84–98. doi:https://doi.org/10.1016/j.trac.2012.09.006.
[40] M.C.U. Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, Chemom. Intell. Lab. Syst. 57 (2001) 65–73. doi:10.1016/S0169-7439(01)00119-8.
[41] A.A. Gomes, M.R. Alcaraz, H.C. Goicoechea, M.C.U. Araújo, The successive projections algorithm for interval selection in trilinear partial least-squares with residual bilinearization, Anal. Chim. Acta. 811 (2014) 13–22. doi:10.1016/j.aca.2013.12.022.
[42] S. Wold, M. Sjöström, L. Eriksson, Partial least squares projections to latent structures (PLS) in chemistry, Encycl. Comput. Chem. 3 (2002).
[43] L. Norgaard, 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. doi:10.1366/0003702001949500.
[44] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Elimination of uninformative variables for multivariate calibration., Anal. Chem. 68

(1996) 3851–3858. doi:10.1021/ac960321m.

[45] D. Broadhurst, J.J. Rowlandb, D.B. Kelp, Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression , with applications to pyrolysis mass spectrometry, Anal. Chim. Acta. 348 (1997) 71–86. doi:10.1016/S0003-2670(97)00065-2.

[46] A. Niazi, R. Leardi, Genetic algorithms in chemometrics, J. Chemom. 26 (2012) 345–351. doi:10.1002/cem.2426.

[47] A.S. Barros, D.N. Rutledge, Genetic algorithm applied to the selection of principal components, Chemom. Intell. Lab. Syst. 40 (1998) 65–81. doi:10.1016/S0169-7439(98)00002-1.

[48] R. Leardi, Application of genetic algorithm-PLS for feature selection in spectral data sets, J. Chemom. 14 (2000) 643–655. doi:10.1002/1099-128X(200009/12)14:5/6<643::AID-CEM621>3.0.CO;2-E.

[49] R. Leardi, A. Lupiáñez González, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemom. Intell. Lab. Syst. 41 (1998) 195–207. doi:10.1016/S0169-7439(98)00051-3.

[50] J. V. Roque, W. Cardoso, L.A. Peternelli, R.F. Teofilo, R.F. Teófilo, Comprehensive new approaches for variable selection using ordered predictors selection, Anal. Chim. Acta. (2019). doi:10.1016/j.aca.2019.05.039.

[51] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression, J. Chemom. 23 (2009) 32–48. doi:10.1002/cem.1192.

[52] F. Lindgren, P. Geladi, S. Rännar, S. Wold, Interactive variable selection (IVS) for PLS. Part 1: Theory and algorithms, J. Chemom. 8 (1994) 349–363. doi:10.1002/cem.1180080505.

[53] S. Sæbø, T. Almøy, J. Aarøe, A.H. Aastveit, ST-PLS: a multi-directional nearest shrunken centroid type classifier via PLS, J. Chemom. 22 (2008) 54–62. doi:10.1002/cem.1101.

[54] M. Goodarzi, B. Dejaegher, Y. Vander Heyden, Feature Selection Methods in QSAR Studies, J. AOAC Int. 95 (2012) 636–651. doi:10.5740/jaoacint.SGE_Goodarzi.

[55] Y. Saeys, I. Inza, P. Larranaga, A review of feature selection techniques in bioinformatics, Bioinformatics. 23 (2007) 2507–2517. doi:10.1093/bioinformatics/btm344.

Chapter II

Variable selection for multivariate regression

The contents of this chapter have been adapted from:

Roque JV, Cardoso W, Peternelli LA, Teófilo, RF. Comprehensive new approaches for variable selection using ordered predictors selection. Analytica Chimica Acta 2019; 1075: 57-70.

DOI: /10.1016/j.aca.2019.05.039

Abstract

New strategies of ordered predictors selection (OPS) were developed in this work, making this method more versatile and expanding its worldwide use and applicability. OPS is a recognized method to select variables in multivariate regression and is used by analytical chemists and chemometrists. It shows high ability to improve the prediction of models after the selection of a few and important variables. At the core of OPS is sorting variables from informative vectors and systematically investigating the regression models to identify the most relevant set of variables by comparing the cross-validation parameters of the models. Nevertheless, the first version of the OPS method performs variable selection using only one informative vector at a time and is limited to just one variable selection run. Then, three new strategies were proposed. First, an automatic method was developed to perform variable selection using several informative vectors and their combinations. Second, the feedback OPS is presented, in this new strategy the pre-selected variables would return to a new selection. Last, a method to apply OPS in full array subdivisions called OPS intervals was established. Initially, the new strategies were applied in the six datasets used in the original OPS paper to compare the prediction performance with the new OPS algorithms. After that, twelve new datasets were used to test and compare the new OPS approaches with other variable selection methods, genetic algorithm (GA), the interval successive projections algorithm for PLS (*i*SPA), and recursive weighted partial least squares (rPLS). The new OPS approaches outperformed the first OPS version and the other variable selection methods. Results showed that in addition to greater predictive capacity, the accuracy in the selection of expected variables is highly superior with the new OPS approaches. Overall, the new OPS provided the best set of selected variables to build more predictive and interpretative regression models, proving to be efficient for variable selection in different types of datasets.

Keywords: Multivariate regression; feature selection; chemometrics; informative vector; prediction power.

Introduction

Multivariate regression models describe the relationship between the dependent and independent variables when more than one measurement is acquired for each sample [1,2]. There are several multivariate regression methods, such as multiple linear regression (MLR), support vector machine regression (SVMR) and partial least squares regression (PLS). However, the latter is the most commonly used for building first order inverse regression models from data of chemical origin [3].

PLS regression can deal with a large number of highly correlated independent variables, band overlaps, and experimental noise [4,5]. Also, it enables simultaneous modeling and the prediction of more than one analyte [2,3]. However, in some situations, the models cannot provide a satisfactory prediction. This may occur because not all variables are equally essential and variations in the concentrations of the chemical components of interest present in the sample do not cause the same change in all variables [6]. Also, there are non-linear and low signal-to-noise ratio regions that might affect the model. This evidence indicates that the proper selection of variables can significantly improve the efficiency of the multivariate regression model, in addition to making it simpler for interpretations [1,6].

Variable selection is a significant step in multivariate regression, and it has become a fundamental tool in many different research areas. This is because of increased database dimensions, meaning that some variables may be redundant, irrelevant or represent noise [2,6,7]. Models built after the removal of non-informative variables will produce better predictions, a better interpretation with the selection of markers or biomarkers, lower measurement costs through the use of portable instruments, and a decrease in data processing time [1,6].

Several feature-selection methods [8–11] have been applied in several works in the literature [12–16]. Nevertheless, most selection methods are not generalized or efficient for all types of datasets. Different analytical techniques provide distinctive predictor types, and each technique has some well-defined peculiarities [17]. In 2009, Teófilo et al. presented first version of the ordered predictors selection (OPS) algorithm [17], with the advantage of being simple, fast and efficient for the selection of any variable type. Since the publication in 2009, the OPS method has had its potential recognized in the literature [7,18,19], showing a high ability to improve the prediction of multivariate regression models with few variables [15,20,21]. However, the original OPS algorithm performs the variable selection using only one

informative vector at a time and is limited to just one variable selection run, which turn out to be limitations of the original version when searching for the best set of variables.

Therefore, the aim of this work was to make a more versatile OPS, expanding its worldwide use and applicability. The development of three new OPS approaches was proposed: *i*) automatic OPS (a*uto*OPS), which automatically performs all calculations using either or both informative vectors and its combinations, so the best one is chosen; *ii*) feedback OPS (*feed*OPS), wherein the pre-selected variables would go through a new selection run; and *iii*) interval OPS (*i*OPS), the option to apply OPS in subdivisions of the full array. The new OPS algorithms were developed to be easily understood and executed. Besides that, the algorithms were applied to several types of dataset (sparse or non-sparse), with the proposal to select interpretive variables with greater predictability, and high reproducibility, *i.e.*, selecting the same set of variables when executing the selection more than once.

Theory Background

Model evaluation

Regression models were evaluated using statistical parameters such as the root mean square error (*RMSE*) and correlation coefficient (*R*), according to equation (II.1) and equation (II.2), respectively.

$$RMSE = \sqrt{\sum_{i}^{I_m} (y_i - \hat{y}_i)^2} \Big/ I_m \qquad \textbf{(II.1)}$$

$$R = \sum_{i=1}^{I_m} (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y}) \Big/ \sqrt{\sum_{i=1}^{I_m} (\hat{y}_i - \bar{\hat{y}})^2} \sqrt{\sum_{i=1}^{I_m} (y_i - \bar{y})^2} \qquad \textbf{(II.2)}$$

where y_i and $\hat{y}_i$ are measured and predicted values for each *i* sample, respectively. The $\bar{y}$ is the **y** values average. When calibration is used, I_m represents the number of samples in the calibration set, and the error and correlation coefficient are the root mean square error of calibration (*RMSEC*) and the correlation coefficient of calibration (R_c), respectively. When internal cross-validation (CV) is used, I_m represents the number of samples in the cross-validation set, and the error and correlation coefficient are the root mean square error of cross-validation (*RMSECV*) and the correlation coefficient of cross-validation (R_{CV}), respectively. When external validation is used, I_m represents the number of predicted samples (*P*) and, in this

instance, the error and correlation coefficient are the root mean square error of prediction (*RMSEP*) and the correlation coefficient of prediction (R_P), respectively.

OPS theory

The OPS method is based on obtaining an informative vector containing information about the best independent variables for prediction. This informative vector can be obtained of several ways from calculations performed with **X** matrix columns (independent variables) and dependent variables (**y**), and its length is equal to the number of independent variables. The original OPS method calculates informative vectors using the regression coefficients (REG), the correlation between each column of matrix **X** with **y** (COR), residual information of the reconstructed matrix with *h* latent variables (SQR), variable importance on projection (VIP), net analyte signal (NAS), and covariance procedures (COV). Details about these vector calculations can be found elsewhere [17].

After obtaining an informative vector, the independent variables of **X** matrix are differentiated according to their corresponding absolute values in the informative vector. The highest absolute value corresponds to the more important independent variable, and the differentiated variables are sorted in descending order of absolute values magnitude. Multivariate regression models are built and evaluated using cross-validation approach. An initial subset of variables (window) is defined to build and evaluate the first model. After, this initial subset is extended by the addition of a fixed number of variables (increment) over the window, and a new model is built and evaluated. Further increments are added until all or a percentage of variables are taken into account, and cross-validation parameters are calculated for each model. Lastly, the variable subsets are compared using the quality parameters calculated during validations, and the best variable subset is defined [17]. Figure II.1 shows a general scheme for the original OPS algorithm.

Core algorithm

The core algorithm of the OPS method consists of the following steps.

calculate an informative vector;
sort in descending order the absolute values of the informative vector and store the sorting index;
sort **X** variables considering the previous sorting index;
define an initial number of variables (window) to be investigated in the ordered **X** array;
define a fixed number of variables to be added over the window (increment);

build multivariate regression models starting with the initial window and then its extension by addition of increments;

store the index of the variables investigated in each subset (window plus increment);

store cross-validation parameters of each model built by each subset;

choose the best set of variables based on cross-validation parameters.

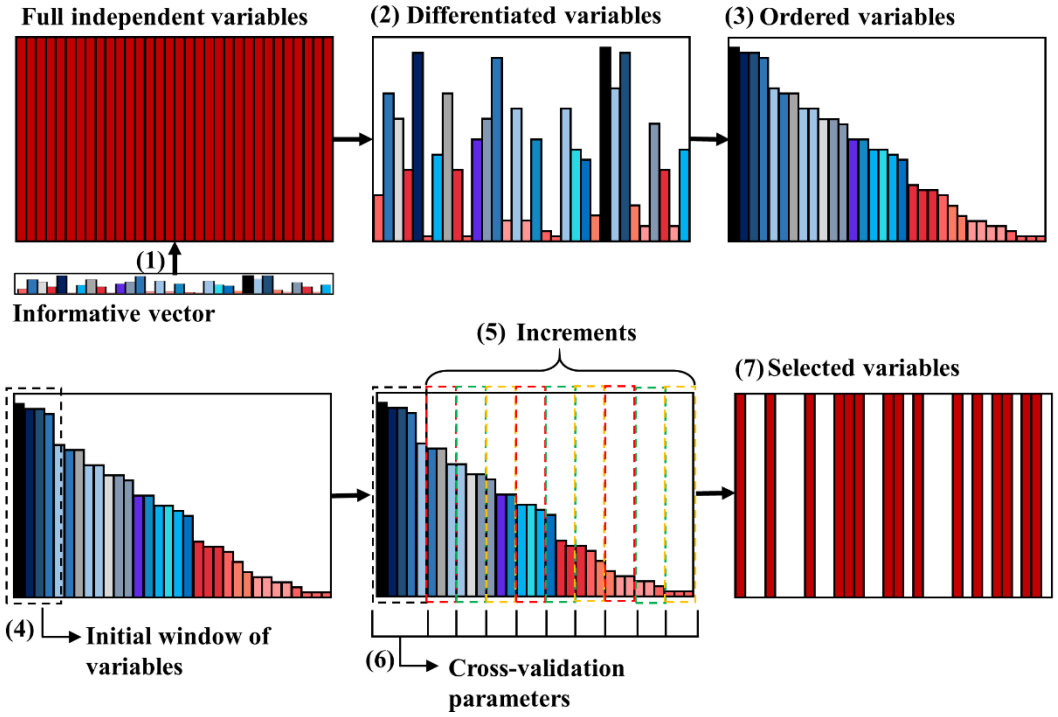

Figure II.1 General scheme of variable selection steps using OPS algorithm. (1) Informative vectors are obtained; (2) the data matrix is differentiated in according to the corresponding absolute values of the informative vector elements; (3) the differentiated variables are sorted in descending order; (4) an initial subset of variables (window) is defined to build and evaluate the first model; (5) this initial subset is extended by the addition of a fix number of variables (increment) over the window, and further increments are added until all, or some percentage of variables are taken into account; (6) cross-validation parameters are calculated for each model; (7) the variable subsets are compared using the quality parameters calculated during validations, and the best variables subset is defined. The different colors represent the weight of the variables and were randomly assigned.

Automatic OPS (autoOPS)

The choice of informative vector is a critical step in the OPS method. Then, the automatic OPS was proposed to perform all steps of variable selection using several informative vectors and provided the best results. In the new approach *auto*OPS, beyond the six informative vectors calculated in the original OPS, two new vectors were used, being the univariate regression between each column of matrix **X** with **y** (URXY), and the vector of weights (WGHT) obtained

by the NIPALS [22] (nonlinear iterative partial least squares) algorithm. The URXY vector contains the information about the regression coefficient obtained by each *j* column of **X** ($\mathbf{x}_j$) with **y** response and the difference between the measured and predicted **y**. This vector can be calculated as follows:

$$\begin{aligned}
&\hat{\mathbf{b}}_j = (\mathbf{x}_j^t \mathbf{x}_j)^{-1} \mathbf{x}_j^t \mathbf{y} \\
&\hat{\mathbf{y}}_j = \mathbf{x}_j \hat{\mathbf{b}}_j \\
&\mathbf{ey}_j = \mathbf{y} - \hat{\mathbf{y}}_j \\
&\mathbf{urxy}_j = \hat{b}_j(2) / (\mathbf{ey}_j^t \mathbf{ey}_j)
\end{aligned} \qquad \textbf{(II.3)}$$

A univariate regression (**y** = **xb**) is performed using each predictor $\mathbf{x}_j$ and **y**. The regression coefficients **b** (a vector with two elements, being the intercept and slope, respectively) are obtained by inverse least squares. The residue ($\mathbf{ey}_j$) between the predicted $\hat{\mathbf{y}}$ and **y** is calculated, and the URXY value of each variable ($\mathbf{urxy}_j$) is subsequently obtained by ration between the second element of **b** (slope) and the residual sum of squares. A high amount of URXY indicates that the corresponding variable should contain valuable information for the model as the residue for this variable presents a small value.

Besides the individual vectors (eight options, six from the original version of OPS and two new ones), their binary combinations (twenty-eight combinations) and the product of all individual vectors simultaneously (PRODALL) can also be used to search the best set of variables for prediction. All vectors are transformed into absolute values, and the infinite norm was applied. The norm was not applied on COR informative vector. The summary of all informative vectors is shown in Figure II.2.

The new approach performs the selection using one of the following vector input options:

1. Single vector, *i.e.*, only one of the thirty-seven vectors (eight individual, twenty-eight binary combinations and the PRODALL shown in Figure II.2 as the colored boxes).
2. Main vectors, *i.e.*, all eight individual vectors simultaneously (vectors placed on the dashed rectangle in the left side of Figure II.2).
3. Interaction vectors, *i.e.*, all twenty-eight binary combinations of two individual vectors (colored ones above the main diagonal line) plus PRODALL (the top row of Figure II.2) simultaneously.
4. All vectors, *i.e.*, the main and interaction vectors options simultaneously.

In these options, when more than one vector is initialized in the OPS algorithm (options 2, 3, and 4), the correspondent vectors are used. Based on cross-validation parameters, the best vector is chosen to perform variable selection.

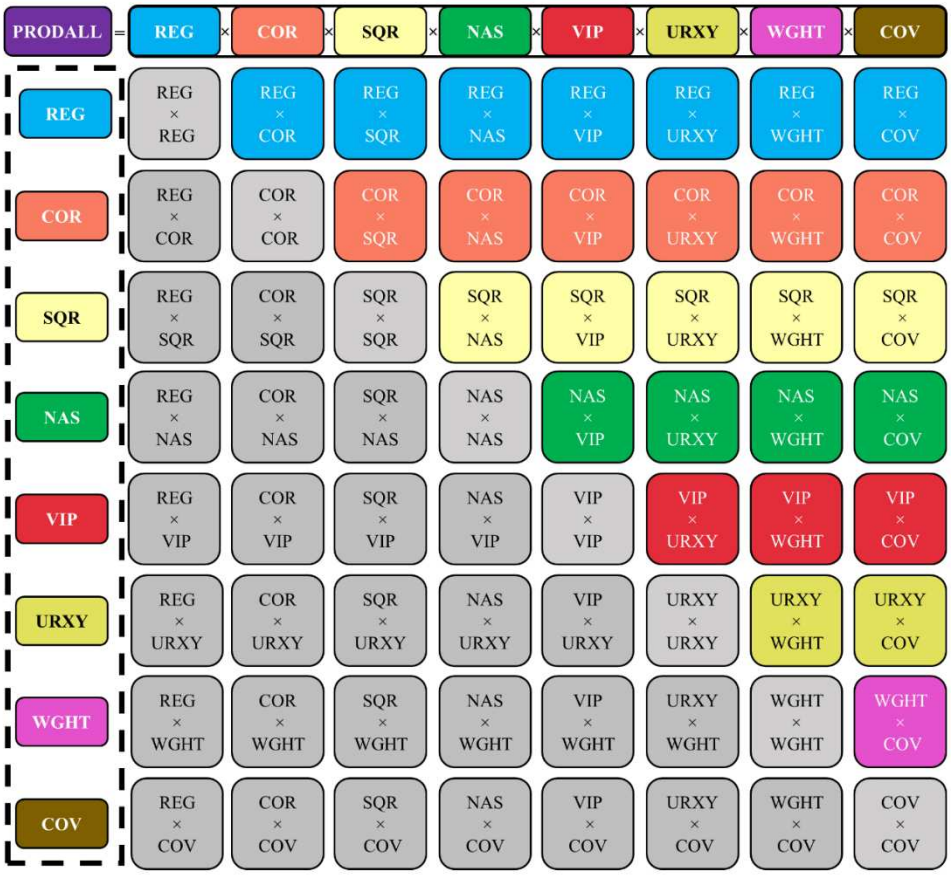

Figure II.2 Summary of all informative vector options to be used in new OPS algorithms for variable selection. REG: regression coefficients; COR: correlation between each column of matrix **X** with **y**; SQR: residual information of the reconstructed matrix with *h* latent variables; NAS: net analyte signal; VIP: variable importance on projection; URXY: univariate regression between each column of matrix **X** with **y**; WGHT: weights; COV: covariance procedures; PRODALL: the product of all single vectors simultaneously.

The aim is to find the best variable subset to build predictive multivariate regression models. Thus, a selection criterion (*sc*) was defined to choose the vector, which is shown in equation (II.4).

$$sc_i = RMSECV_i / R_{CV\,i} \qquad \textbf{(II.4)}$$

where *RMSECV* to R_{CV} ratio of each vector *i*, it is obtained and the vector that shows the lowest *sc* is chosen. Therefore, the set of variables selected is used to build the final model.

Particular attention should be given to the number of latent variables used in OPS algorithms. Firstly, the number of latent variables (*h*Mod) used to build the model was determined based on *RMSECV* values obtained by cross-validation procedure. In addition, some informative vectors, as REG, SQR, VIP, NAS, and WGHT, require a number of latent variables to be obtained. However, sometimes, *h*Mod is not able to generate an informative vector with quality for variable selection. Therefore, to find the best number of latent variables for OPS (*h*OPS), a study using the full dataset is performed by increasing the number of latent variables of the model, starting from the pre-determined *h*Mod, and carrying out the variable selection up to a given number of latent variables. Then, varying the *h*Mod value, the *h*OPS is chosen based on the smallest *RMSECV* value. Hence, two types of latent variables are employed in the OPS algorithm, one representing the number of latent variables for model building (*h*Mod) and the other employed to generate the REG, SQR, VIP, NAS, and WGHT informative vector in OPS method (*h*OPS).

autoOPS algorithm

The algorithm *auto*OPS consists of the following steps.

choose k informative vectors to be studied;
for $i = 1$ to k
  calculate the i informative vector;
  run the **core of OPS algorithm** using the i informative vector;
  store the i set of selected variables;
  apply equation (II.4);
  store sc_i;
end
the lowest *sc* indicates the informative vector that selects the best set of variables.

Feedback OPS (feedOPS)

The *feed*OPS has the purpose of applying feedback in the OPS algorithm, wherein pre-selected variables can return to a new selection run until specific criteria are achieved (Figure II.3). This new strategy works in a loop constrained by one or more rule; when one of them is attained the loop stops.

In *feed*OPS, *RMSECV* was taken as a reference parameter. Convergence is achieved when in two consecutive selection runs, the relative difference (*rd*) (equation (II.5)) is less than an *rd* value defined or when the *RMSECV* value increases instead of decrease. Thus, the previous

loop is taken as the best selection. Besides that, the maximum number of selection runs can be defined. Therefore, the last loop is considered the best selection. In each selection run, the *auto*OPS is applied to provide the best informative vector.

$$rd = RMSECV_n - RMSECV_{n-1} / RMSECV_n \quad \textbf{(II.5)}$$

where n is the number of selection runs.

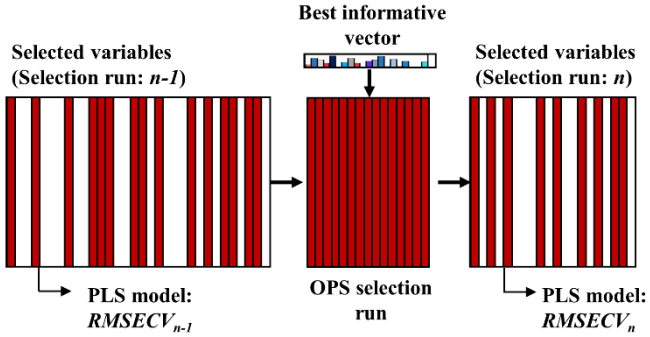

Figure II.3 Scheme of the *feed*OPS algorithm where the pre-selected variables return to a new selection run until specific criteria are achieved. *RMSECV*: root mean square error of cross-validation, *n*: loop counter.

feedOPS algorithm

The algorithm *feed*OPS consists of the following steps.

define convergence criteria (*rd* and *l* = max number of loops);
n = 1;
while $rd_{calc} > rd$ *or* $n < l$
  run ***auto*OPS** algorithm;
  store $RMSECV_n$ of the selected informative vector*;*
  % The *RMSECV* of full **X** matrix was used to compare with *RMSECV* in the first loop.
  calculate rd_{calc} (equation (II.5));
    if $RMSECV_n - RMSECV_{n-1} > 0$
      break.
    end
  $n = n + 1$;
end
obtaining the final set of selected variables.

OPS Intervals (iOPS)

In this new approach, the OPS is applied in subdivisions of the full array. This search strategy is called OPS intervals (*i*OPS) and is showed in Figure II.4.

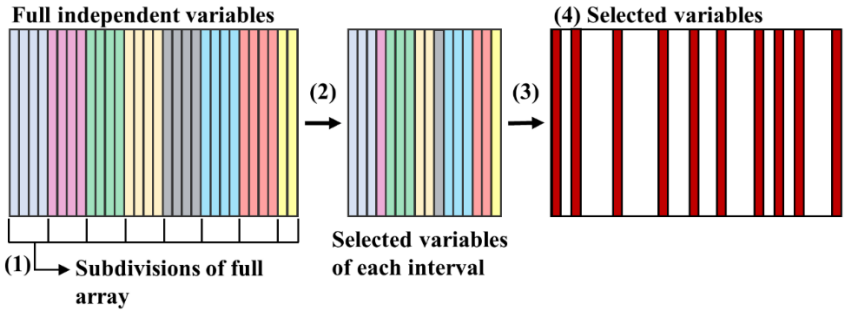

Figure II.4 Scheme of *i*OPS algorithm where the variable selection is performed in each interval; subsequently, these selected variables are ordered, and a new selection is performed to find the best set of variables. (1) The array is subdivided; (2) *auto*OPS or *feed*OPS is applied in each interval; and a new matrix is created with the selected variables in each interval (3) *auto*OPS or *feed*OPS is applied in the new matrix; (4) the best set of variables is reached.

Firstly, the number of variables in each interval is defined. The entire array is subdivided, and the variable selection is performed. In each interval, the *auto*OPS or the *feed*OPS is applied. This step is mainly used to reduce the number of variables in each interval. The intervals comprising only the selected variables will be merged into a new matrix and a new variable selection will take place considering predetermined window and increment. The *auto*OPS (or *feed*OPS) is applied in the new array, and the best set of variables is reached.

iOPS algorithm

The algorithm *i*OPS consists of the following steps.

define the number of variables in each interval (at least 50);
apply ***auto*OPS** or ***feed*OPS** algorithm in each interval;
find the best set of variables for each interval;
create a new array containing the selected variables in each interval;
apply ***auto*OPS** or ***feed*OPS** algorithm in the new array;
obtaining the final set of selected variables.
% We nickname this algorithm of *crème de la crème*.

Once all new approaches have been explained, Figure II.5 shows a chart that represents the options to apply the new OPS algorithms and summarizes the names of different options of OPS algorithm application.

Each new approach combined with each vector option generates sixteen ways to apply the OPS. For example, using *auto*OPS with a single vector, the option is called autoS. The *feed*OPS with the main vector option is named *feed*M. The application of *i*OPS with *auto*OPS and all vectors option is called *i*autoA.

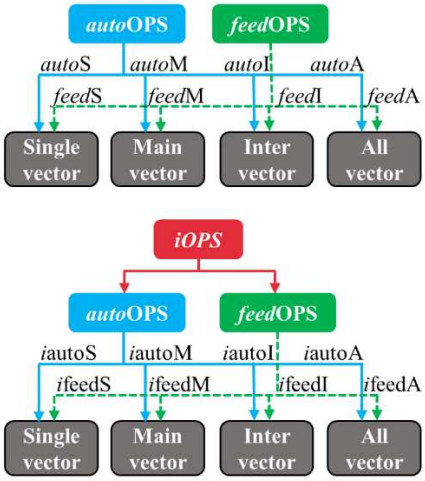

Figure II.5 Scheme of the new OPS algorithms (*auto*OPS, *feed*OPS, and *i*OPS) combined with the four vector options. Names of OPS options derived by the combination of new OPS approaches with all vector options are arranged in the lines for the arrows.

Experimental

The OPS algorithms were written in function .m to MATLAB 2019a (Math Works, Natick, USA) and are fully authorial (available at www.deq.ufv.br/chemometrics). This section is about the validation of the new approaches of OPS (*auto*OPS, *feed*OPS, and *i*OPS). Initially, the new approaches were applied in the six datasets used in the original OPS paper [17] to compare the prediction performance with the new OPS algorithms. Then, twelve new datasets, comprising different types of data, were used to compare the performance of the new OPS approaches with other variable selection algorithms used by chemometrists, *i.e.*, genetic algorithm [10] (GA), intervals successive projections algorithm for PLS [8] (*i*SPA), and recursive weighted partial least squares [11] (rPLS). All calculations were performed in MATLAB environment.

Modeling

The six datasets studied in the original OPS paper were used to build PLS models using variables selected by the new OPS approaches (*auto*OPS, *feed*OPS, and *i*OPS). They were tested using all vectors, the most complete OPS option. The calibration and prediction sets were the same used in the original paper [17]. In addition, the method leave-*N*-out cross-validation

was applied, where *N* was set as 10% of the total sample number in the training set. The range calibrated, the size of calibration and prediction sets, and the pre-treatment performed in each dataset can be found elsewhere [17]. The new OPS algorithms were applied using different windows and increments of variables for each dataset, and 100% of variables were tested. In *feed*OPS, a minimum difference of 2% between two consecutive *RMSECVs* and a maximum number of ten loops were set as convergence criteria. In *i*OPS, when the *feed*OPS option was used to run the selection, the convergence criteria were the same as used in *feed*OPS. The full matrix was divided into intervals of 10% of the its size, limited in at least fifty variables. Additionally, *h*OPS was calculated for each interval in *i*OPS.

Twelve new datasets were used to build PLS models using all variables of the **X** matrix (full models) and selected variables using the new OPS, GA, *i*SPA, and rPLS algorithms. All three new approaches of OPS were tested for each dataset using all vectors, the most complete OPS option. A total of thirty-seven informative vectors were evaluated. In order to compare the OPS with the other three variable selection algorithms, the new strategy that provided the best OPS model for each dataset was chosen between the three new strategies available using the all vectors option. This choice was made based on the parameters described in the section "*Theory Background – Model evaluation"*.

The full datasets were split into calibration and prediction sets using Kennard-Stone algorithm [23]. Table II.1 shows information about the property and range calibrated, the size of calibration and prediction sets, and the pre-treatment performed in each dataset. The **y** variable was mean centered for all the properties and different pre-treatments and their combinations were studied for each full matrix **X**. The pre-treatments tested were mean center, autoscale, smoothing, first and second derivative, multiplicative scatter correction (MSC), normalization, baseline, and standard normal variate (SNV). The combinations of two, three and four pre-treatments were also studied. The pre-treatments that presented a model with the lowest *RMSECV* value was chosen for each dataset. All algorithms performed the variable selection with datasets equally pre-treated.

The new OPS algorithms were applied using the window of 10 and increments of 5 variables, 100% of variables were tested, random cross-validation was applied, where splits were set at 10% of **X** matrix rows. Only QSAR models were built using leave-one-out cross-validation. In *feed*OPS, as convergence criteria, 2% as the minimum difference between two consecutive *RMSECV*s and ten as the maximum number of loops.

Table II.1 Information about datasets used to perform validation of new strategies of OPS algorithm.

Dataset	**Voltammetry**	**Fluorescence**	**MS**	**Raman**	**NIR**	**UV**
Property	Ascorbic Acid	Catechol	Ethanol	Iodine Value	Lignin	Phorbol esters
Range	21.5 - 100.0	1.0 - 8.0	12.79 - 15.09	52.0 - 69.0	18.04 - 28.36	0.72 - 3.58
Unit	µmol L^{-1}	µmol L^{-1}	vol (%)	g I_2 100 g fat $^{-1}$	% (w/w)	mg g^{-1}
Size	22 × 525 (Cal) 10 × 525 (Pred)	28 × 221 (Cal) 10 × 221 (Pred)	34 × 200 (Cal) 10 × 200 (Pred)	75 × 5667 (Cal) 30 × 5667 (Pred)	216 × 1038 (Cal) 40 × 1038 (Pred)	106 × 656 (Cal) 32 × 656 (Pred)
Column-wise pre-treatment	Mean Center	Mean Center	Mean Center	Auto	Mean Center	Mean Center
Row-wise pre-treatment	Baseline	None	None	SNV	Baseline + 2nd Deriv (3) + MSC	1st Deriv (15)
Dataset	**NMR**	**GC**	**Vis/NIR**	**XRF**	**QSAR**	**MS**
Property	Pentanol	Overall Quality	Xylene	Ni	MMP-2 pIC_{50}	Peroxide Value
Range	0 - 100	1.13 - 4.50	4.00 - 15.01	0.062 - 15.890	6.25 - 8.28	1.05 - 33.62
Unit	%	-	weight percent (%)	%	-	meq Kg^{-1}
Size	181 × 2334 (Cal) 50 × 2334 (Pred)	119 × 2294 (Cal) 40 × 2294 (Pred)	25 × 316 (Cal) 5 × 316 (Pred)	12 × 261 (Cal) 3 × 261 (Pred)	25 × 439 (Cal) 6 × 439 (Pred)	20 × 106 (Cal) 5 × 106 (Pred)
Column-wise pre-treatment	Mean Center	Auto	Mean Center	Mean Center	Auto	Mean Center
Row-wise pre-treatment	None	MSC + 1st Deriv (19) + Norm	1st Deriv (3)	None	None	None

Cal: calibration set; Pred: prediction set; Auto: autoscaling; SNV: standard normal variate; 1st or 2nd Deriv: type of derivative with the window between parenthesis; MSC: multiplicative scatter correction; Norm: normalize to unit area; MMP-2 pIC_{50}: Half-maximal inhibitory concentration negative logarithm of type 2 matrix metalloproteinases.

In *i*OPS, when the option to run the selection using *feed*OPS was used, the convergence criteria were the same as those used in *feed*OPS. The full matrix was divided into intervals of 10% of the its size, limited in at least fifty variables. Additionally, *h*OPS was calculated for each interval in *i*OPS. Random cross-validation was applied to build full models, where splits were set at 10% of **X** matrix rows. Although the OPS algorithms require several parameters to optimize the variable selection, only **X** and **y** are mandatory inputs for all new OPS algorithms. It is possible to perform the variable selection using some default parameters and automatic choices, but the result may not be the optimal one.

The GA is a method used for solving optimization problems based on natural selection processes and genetics that mimic biological evolution [10]. Initially, a starting population is constituted for a series of different individuals and the number of individuals is defined as the population with a defined size. These individuals are analyzed and crossed according to the parameters introduced into the algorithm. At the end of the process, a vector consisting of zeros and ones indicates whether variables should be included (1) or not (0). The parameters used to perform the variable selection were optimized (results not shown) and the following conditions were used: a population of 52, a maximum generation of 300, a mutation rate of 0.008, a window width of 1, convergence of 80, 50 terms included at initiation, cross-over rule of 2, random cross-validation was applied with the number of subsets to divide data into for cross-validation of 10, cross-validation iteration of 1 and 3 replicate runs. The GA variable selection was performed using the *.m* functions from PLS-Toolbox 8.2 (Eigenvector Research, Inc. Wenatchee, USA).

The *i*SPA combines the noise-reduction properties of PLS with the possibility of discarding non-informative variables with SPA-MLR algorithm [24]. In this method, it is assumed that the variables have been divided into non-overlapping intervals, usually, with the same length. First, the columns of **X** are partitioned according to the intervals of variables previously defined and subjected to a sequence of projection operations that result in the creation of chains. Second, PLS models are built using leave-one-out cross-validation for each combination of intervals, and the best combination of intervals is then chosen by the smallest *RMSECV* [8]. The *i*SPA algorithm was applied dividing the spectra into 20 intervals and the maximum number of intervals was selected as 20. This algorithm for MATLAB was provided by the authors.

The rPLS method uses the dependent variable **y** to guide the variable weighting recursively. This method iteratively reweights the variables using the regression vector

calculated by PLS [11]. Random cross-validation with splits of 10 and the number of iterations infinite were the default settings used in variable selection. This algorithm for MATLAB was retrieved from www.models.life.ku.dk/algorithms.

The variable selection was carried out in triplicate on the calibration set. Thus, three models were built for OPS, GA, *i*SPA, and rPLS. Each model was applied on the prediction set, and the calculated *RMSEP* values were pairwise compared by Tukey test. Differences with $p < 0.05$ were considered significant.

Datasets

The six datasets used in the original OPS paper were near-infrared spectroscopy (NIR), fluorescence spectroscopy, a simulated dataset, Raman spectroscopy, gas chromatography (GC), and quantitative structure-activity relationship (QSAR) data.

For NIR spectra of diesel samples [25], the following physical properties were modeled: boiling point at 50% recovery, cetane number, density, freezing temperature of the fuel, total aromatics, and viscosity. Mixtures of standard solutions of six different analytes were used in fluorescence [26]: catechol, hydroquinone, indole, resorcinol, *L*-tryptophane, and *DL*-tyrosine. The simulated dataset consisted of twenty mixtures simulated by using ultraviolet-type spectra from four analytes and their respective concentrations randomly generated. For Raman spectra of *Escitalopram*® tablets [27], the dependent variable referred to the amount of active substance in the tablets. GC dataset of fuel samples (Pirouette® software – Infometrix, Inc) is formed by thirty-five independent variables consisting of the chromatograms peak areas and three dependent variables comprising the following physical properties: flash point, freeze point, and specific gravity. The QSAR dataset [28] consists of fourteen molecular descriptors for forty-eight HIV-1 protease inhibitors and the dependent variable was the *in vitro* inhibition activity.

For NIR dataset, the new OPS algorithms were applied using window of five and increments of two variables (window 5, increment 2), Raman dataset (window 10, increment 5), fluorescence dataset (window 5, increment 2), GC dataset (window 2, increment 1), QSAR dataset (window 2, increment 1), and simulated dataset (window 5, increment 1).

The twelve new datasets used to compare full models and selected variables using the new OPS, GA, *i*SPA, and rPLS algorithms are described below.

Voltammetry: This dataset was presented by Teófilo [29] and consists of a mixture of standard solutions of three biological interest compounds (dopamine, uric acid, and ascorbic

acid). Square wave voltammetry (SWV) with boron-doped diamond electrode was employed to quantify these analytes. The potential range investigated was from 0.01 to 1.5 V.

Fluorescence Spectroscopy: This dataset was presented by Teófilo [29] and consists of a mixture of six standard phenol solutions (hydroquinone, guaiacol, p-cresol, m-cresol, catechol, and phenol). The excitation wavelength was fixed at 275 nm, and the emission wavelength was scanned from 270 to 380 nm.

Mass Spectrometry (MS): (1) The first mass dataset was obtained from http://www.models.ku.dk/datasets and was presented by Skov et al. [30]. A mass spectrum was obtained for each sample of wine in the range of 5 – 204 m/z using the electron ionization mode at 70 eV. Fourteen properties were evaluated in the original work [30], but only ethanol was calibrated in this work. (2) The second contains mass spectra collected from the headspace of butter samples that had been artificially aged to produce various levels of spoilage. The reference values for rancidity are peroxide values. The data were supplied by Leatherhead, UK, and obtained from Pirouette® (Infometrix, Inc) software. The mass spectrum was obtained in the range from 44 – 149 m/z.

Raman Spectroscopy: This dataset was presented by Lyndgaard et al. [31], and it is available at http://www.models.ku.dk/datasets. The samples were 16 pork carcasses taken from the daily production stock of a slaughterhouse for determining the fatty acid composition of pork backfat as a function of the iodine depth profile. The Raman spectra were acquired using a total of 16 accumulations of 1 s exposure and were stored as Raman shifts in the range 1800 – 200 cm^{-1}.

Near Infrared Spectroscopy (NIR): This dataset is composed of sugarcane leaf NIR spectra that was used to predict the lignin content of sugarcane stalks, and it was provided by Assis et al. [16]. The NIR spectra were acquired from 10000 to 4000 cm^{-1}.

Ultraviolet Spectroscopy (UV): This dataset was obtained from Roque et al. [32], and consists of UV spectra of *Jatropha curcas* extracts in the range of 210 to 350 nm. Phorbol esters content was the modeled property.

Nuclear Magnetic Resonance Spectroscopy (NMR): This dataset is available at http://www.models.ku.dk/datasets, and was presented by Winning et al. [33]. It consists of NMR spectra of a designed set of 231 simple alcohol mixture (propanol, butanol, and pentanol), and each spectrum was acquired in the range of 0.64 to 3.84 ppm.

Gas Chromatography (GC): Chromatographic profiles of volatile roasted coffee compounds provided by Ribeiro et al. [34] to predict scores of coffee beverage overall quality.

The chromatograms dataset used in this work comprises of the range from 1.25 to 21.83 min of retention time.

Visible and Near Infrared Spectroscopy (Vis/NIR): The data were obtained from Pirouette® (Infometrix, Inc) software. This dataset contains spectra of hydrocarbon mixtures (heptane, isooctane, toluene, xylene, and decane) from two different diode array spectrometers. Absorbances from 470 to 1100 nm were collected.

X-Ray Fluorescence Spectrometry (XRF): X-ray fluorescence spectra were obtained from Pirouette® (Infometrix, Inc) software. Wang et al.[35] presented this dataset. Spectra of nickel alloys plus elemental concentrations in the alloys were measured from $2\theta = 44°$ to 70°. The nickel concentrations were determined by wet chemistry.

Quantitative structure-activity relationship (QSAR): De Melo [36] presented this dataset of matrix metalloproteinases type 2 (MMP-2) with 31 cinnamoyl pyrrolidine derivatives, where 439 molecular descriptors were obtained.

Results and Discussion

Algorithms

The finest OPS model of each dataset is presented in Table II.2; regarding this, no tendency is observed. In order to present the detailed results obtained using each one of the new approaches (*auto*OPS, *feed*OPS, and *i*OPS), it was selected a dataset that shows the best result using each of one them. For *auto*OPS was chosen XRF; for *feed*OPS, voltammetry; and for *i*OPS, GC dataset.

The best OPS model obtained for XRF dataset was using the *auto*OPS algorithm. In Figure II.6A1 the *sc* value obtained for each of the thirty-eight vectors is shown. It is noticeable that the vector SQR × URXY showed the lowest value (dashed red line). Additionally, not all vectors were able to improve prediction quality regarding the full model (solid red line). Some of them presented consistent improvement of model prediction. These results indicated that vector combinations performed better than individual vectors. This conclusion cannot be generalized for all datasets, but it is enough to show the importance of combinations. Figure II.6A2 shows the OPS plot for recognition of the best set of variables. It is a detail of the best result shown in Figure II.6A1 for SQR × URXY vector. This plot displays an increase of the variables number, related to the addition of new variables to the first variables window, and a decrease in *RMSECV* and increase in R_{CV}. This type of plot is very useful to visualize the

variable selection and help in choosing the best set. The ideal subset of variables is found when the ration of *RMSECV* and R_{CV} is minimum (equation (II.4)).

Table II.2 Best OPS model, informative vector and the number of latent variables used to perform the selection (*h*OPS) obtained for each dataset.

Dataset	Property	OPS† model	Vector
Voltammetry	Ascorbic Acid	*feed*	NAS (19)
Fluorescence	Catechol	*i*auto	SQR (5) × COV
MS	Ethanol	*i*auto	REG (18) × COR
Raman	Iodine Value	*i*auto	URXY
NIR	Lignin	*feed*	REG (19)
UV	Phorbol esters	*auto*	REG(9) × COR
NMR	Pentanol	*auto*	SQR (7)
GC	Overall Quality	*i*feed	REG (16)
Vis/NIR	Xylene	*i*auto	REG (10)
XRF	Ni	*auto*	SQR (11) × URXY
QSAR	MMP-2 pIC_{50}	*i*auto	REG (19)
MS	Peroxide value	*feed*	REG (9) × NAS (9)

MMP-2 pIC_{50}: Half-maximal inhibitory concentration negative logarithm of type 2 matrix metalloproteinases. †Best result. *h*OPS values are presented between parentheses.

Sometimes, an increase of the variables number is accomplished by a small variation in *RMSECV* value. In this way, the subset of selected variables may not be ideal, because instead the *RMSECV* increase with the addition of new noninformative variables, this parameter remained almost constant. Then, an overestimated subset of variables can be erroneously selected.

In Figure II.6A2, with few variables, the OPS model performance is worse than the full model (solid red line), and as the number of variables increases, the *RMSECV* decreases and R_{CV} increases until the optimal number of variables is reached (dashed red line). After that, the *RMSECV* value increases and R_{CV} decreases with the inclusion of noninformative variables.

Voltammetry dataset was used to present the result provided by *feed*OPS (Figure II.6B). For this dataset, eight selection runs were performed. The number of informative variables selected and the *RMSECV* value decreases in each loop until one of the criteria is achieved. The maximum of ten loops and the relative difference between two consecutive *RMSECV* values lower than 2% were set as constraints. Convergence was reached with seven loops, where the difference between *RMSECV* of the seventh and eighth loop was smaller than 2 %. Therefore, the set of variables was selected by the NAS vector in loop 7 (dashed red line). As a result, a

meaningful reduction of the number of variables (from 525 to 14) and improved model performance parameters was reached.

Regarding *i*OPS, the GC dataset results were chosen to be presented. Figure II.6C1 shows the *RMSECV* obtained by each interval of the full array using variable selection. In this example, the selection of variables in each subdivision was not able to enhance the predictive capacity of PLS models. So, in the last step of *i*OPS, the selected variables in each interval were merged into a new matrix. A new variable selection took place considering window and increment of variables, and the best set of variables was obtained.

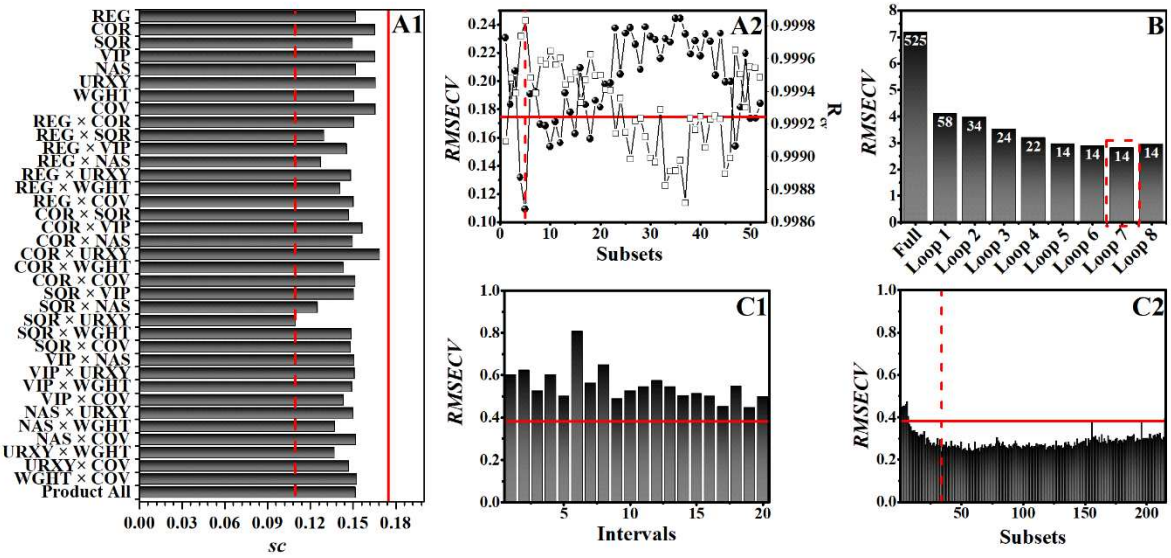

Figure II.6 (A1) Minimum *sc* values obtained for all informative vectors used in the *auto*OPS algorithm for the XRF dataset. The selected vector was detached as dashed red line. (A2) Detailed results of variable subsets obtained by the *auto*OPS algorithm using the SQR × URXY vector and your respective *RMSECV* and R_{CV} values for the XRF dataset. The dashed red line means the selected subset. Black spheres are *RMSECV* values showed in the left axis, and black squares are R_{CV} values showed in the right axis. (B) *Feed*OPS typical plot (number of selection runs and *RMSECV* values) for the voltammetry dataset. The dashed red line indicates the loop where the convergence was attained, and the set of selected variables was obtained. (C1) *i*OPS plot for the GC dataset with *RMSECV* values obtained by PLS models built using the selected variables in each interval of the full array (first variable selection step). (C2) *i*OPS plot for the GC dataset with *RMSECV* values obtained by PLS models built using the subsets of the merged matrix built from selected variables in each interval. A dashed red line shows the subset where the best set of variables were selected. A solid red line represents the *RMSECV* of full model in each subplot of *i*OPS. *sc*: selection criterion; *RMSECV*: root mean square error of cross-validation.

In Figure II.6C2, the result of this last step is shown. *RMSECV* value decreases with the addition of the increments of variables until finding the finest subset of variables (dashed red line) and increases with the inclusion of noninformative variables in the model.

Modeling

The three new OPS approaches provide four different ways to apply OPS since the *i*OPS can be applied using *auto*OPS or *feed*OPS. The PLS models built using the selected variables for the six datasets used in the original OPS paper are shown in Figure II.7.

The number of selected variables is placed in the left-side axis and is represented by bars. The informative vectors are placed inside the bars. The *RMSEP* values are located in the right-side axis and are represented by the black spheres. The *h*Mod for each model is shown below the spheres. In general, the prediction performance of the new approaches outperformed the first version of OPS (OPS*v*1) algorithm in all datasets.

The *h*Mod used in the original paper was taken as reference to run the new OPS algorithms, and for most of the cases, this number was kept and for the others a smaller *h*Mod was chosen automatically. The informative vector used for all datasets in the original paper was REG or some combination with it. With the new approaches, different informative vectors were chosen, including the PRODALL vector for three properties (Figure II.7B, 7H, and 7J), proving the importance of new vectors and their combinations to select more predictive variables. For some datasets, the number of selected variables were slightly higher than the ones selected in the original paper, but the prediction was better.

In Figure II.7R – 7U, it was not possible to apply the *i*OPS algorithm, since the GC (Figure II.7R – 7T) and QSAR (Figure II.7U) datasets have thirty-five and fourteen variables, respectively. For GC (specific gravity – Figure II.7T), the *RMSEP* value was the same for OPS*v*1, *auto,* and *feed* approaches, but the number of selected variables and the informative vector were not the same. For QSAR (Figure II.7U), the *RMSEP* value, the number of selected variables and the informative vector were the same for OPS*v*1, *auto* and *feed* approaches. Both datasets have a few number of variables, and it is possible that this fact influenced these results.

The PLS models obtained using the twelve datasets previously presented are described. The mean results of the performance parameters of PLS models are shown in Table II.3. Concerning the *RMSECV* values, the OPS models were smaller than full models for all datasets. Regarding the cross-validation results, OPS was the most predictive for voltammetry and NIR; GA was the most predictive for fluorescence, MS (ethanol), NMR, Vis/NIR, XRF, QSAR, and MS (peroxide value); rPLS was the most predictive for GC and OPS and rPLS presented similar results for Raman and UV.

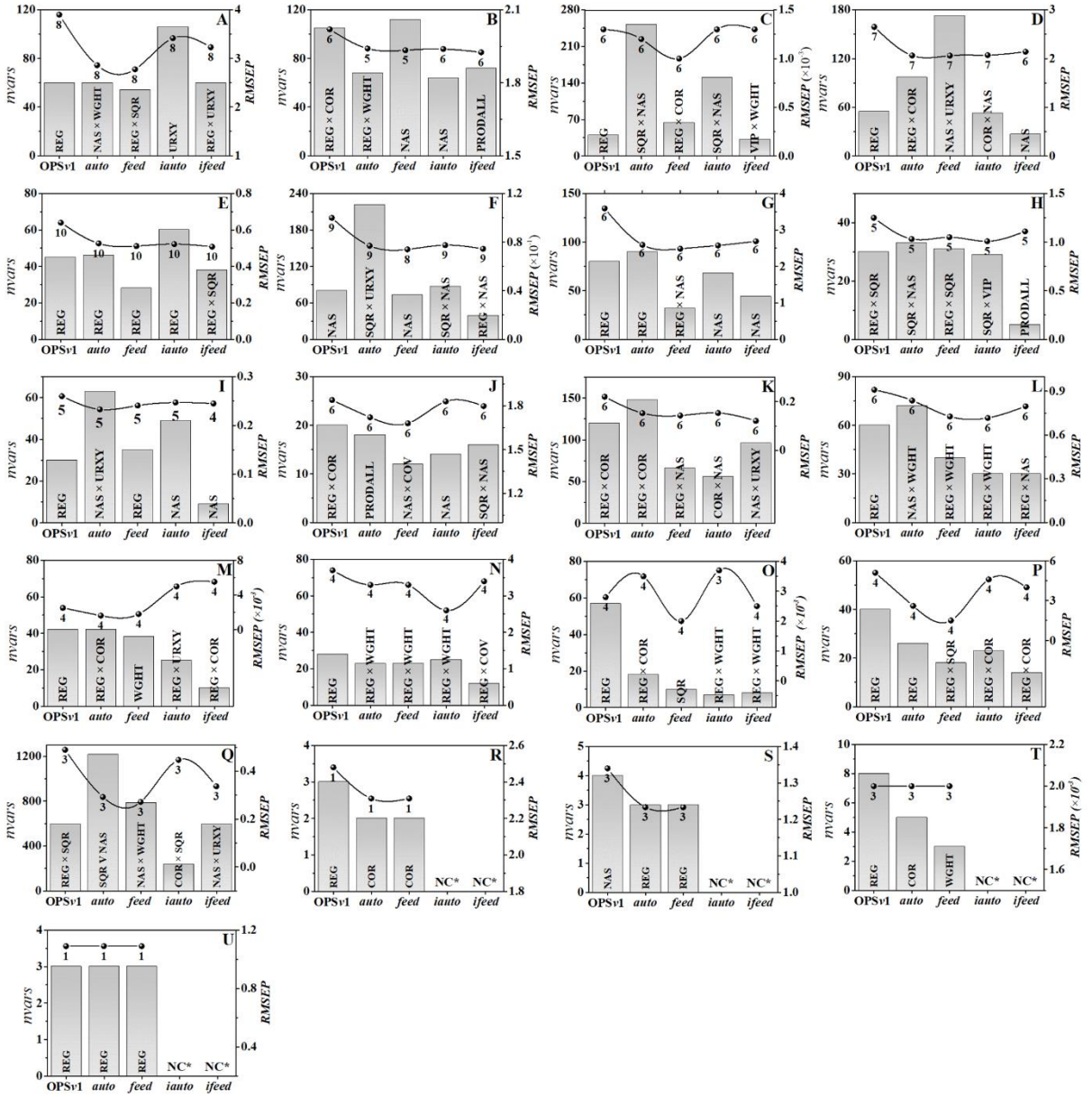

Figure II.7 Number of selected variables placed in the left-side axis (light gray bar) and *RMSEP* values placed in the right-side axis (black spheres). The informative vector and the *h*Mod are placed in bars and black spheres, respectively. (A) NIR – boiling point at 50% recovery, (B) NIR – cetane number, (C) NIR – density, (D) NIR – freezing temperature of the fuel, (E) NIR – total aromatics, (F) NIR – viscosity, (G) Fluorescence – catechol, (H) Fluorescence – hydroquinone, (I) Fluorescence – indole, (J) Fluorescence – resorcinol, (K) Fluorescence – *L*-tryptophane, (L) Fluorescence – *DL*-tyrosine, (M) Simulated – analyte 01, (N) Simulated – analyte 02, (O) Simulated 03 – analyte 03, (P) Simulated – analyte 04, (Q) Raman – active substance in *Escitalopram* ® tablets, (R) GC – flash point, (S) GC – freeze point, (T) GC – specific gravity, and (U) QSAR – *in vitro* inhibition activity. *RMSEP*: root mean square error of prediction, *h*Mod: number of latent variables of the model. *NC: not calculated.

Table II.3 Performance parameters of PLS models obtained using all variables and those selected by OPS, GA, *i*SPA, and rPLS algorithms.

	Voltammetry - Ascorbic Acid					**Fluorescence - Catechol**				
	Full	**OPS**	**GA**	***i*SPA**	**rPLS**	**Full**	**OPS**	**GA**	***i*SPA**	**rPLS**
nvars	525	14	74 – 80	52 – 79	7	221	30	32 – 35	33 – 166	28
*h*Mod			6			3				
RMSECV	8.2	3.4	4.8	10	3.7	0.6	0.4	0.3	0.6	0.5
R_{CV}	0.943	0.991	0.982	0.92	0.989	0.989	0.994	0.997	0.989	0.993
RMSEP	7.6^{c}	4.0^{a}	6.1^{b}	9.2^{d}	8.4^{c}	0.8^{b}	0.7^{a}	0.6^{a}	0.9^{b}	0.8^{b}
R_P	0.969	0.995	0.983	0.962	0.963	0.989	0.994	0.994	0.988	0.99
	MS - Ethanol					**Raman - Iodine Value**				
	Full	**OPS**	**GA**	***i*SPA**	**rPLS**	**Full**	**OPS**	**GA**	***i*SPA**	**rPLS**
nvars	112†	10	13 – 22	79	7	5667	25	2132 – 2224	284 – 3116	84
*h*Mod			6			2				
RMSECV	0.46	0.36	0.29	0.48	0.35	2.1	2	2.1	2.5	2
R_{CV}	0.16	0.666	0.686	0.1	0.517	0.763	0.801	0.785	0.644	0.806
RMSEP	0.68^{e}	0.36^{a}	0.42^{b}	0.48^{c}	0.56^{d}	2.0^{a}	1.9^{a}	2.0^{a}	3.2^{b}	2.4^{ab}
R_P	0.22	0.865	0.802	0.783	0.551	0.88	0.908	0.892	0.72	0.848

nvars: number of variables; *h*Mod: number of latent variables of the model; *RMSECV*: root mean square error of cross-validation; R_{CV}: correlation coefficient of cross-validation; *RMSEP*: root mean square error of prediction, R_P: correlation coefficient of prediction. Units: Voltammetry – ascorbic acid (μmol L^{-1}), Fluorescence – catechol (μmol L^{-1}), MS – ethanol (vol - %), Raman – iodine value (g I_2 100g fat^{-1}).†Full dataset after removal of variables that not present variance. Different superscript letters mean that the models are significantly different.

Table II.3. (continued).

	NIR - Lignin					UV - Phorbol Esters				
	Full	**OPS**	**GA**	***i*SPA**	**rPLS**	**Full**	**OPS**	**GA**	***i*SPA**	**rPLS**
nvars	1038	135	293 – 358	518 – 882	149	656	50	148 – 164	66	51
*h*Mod			10					8		
RMSECV	1.42	0.59	0.81	1.53	0.88	0.29	0.23	0.3	0.25	0.23
R_{CV}	0.733	0.959	0.921	0.698	0.906	0.855	0.912	0.849	0.89	0.909
RMSEP	0.89^{c}	0.67^{a}	0.75^{b}	1.02^{d}	0.80^{b}	0.27^{c}	0.24^{a}	0.28^{d}	0.26^{b}	0.32^{e}
R_P	0.928	0.96	0.948	0.902	0.941	0.947	0.958	0.943	0.952	0.924
	NMR - Pentanol					**GC - Overall Quality**				
	Full	**OPS**	**GA**	***i*SPA**	**rPLS**	**Full**	**OPS**	**GA**	***i*SPA**	**rPLS**
nvars	2334	198	985 – 1167	2101	199	2294	185	762 – 772	115 – 1492	736
*h*Mod			4					5		
RMSECV	0.68	0.65	0.61	0.68	0.69	0.37	0.28	0.2	0.49	0.19
R_{CV}	0.999	0.999	0.999	0.999	0.999	0.918	0.953	0.977	0.847	0.979
RMSEP	2.45^{b}	2.32^{a}	2.45^{b}	2.45^{b}	2.44^{b}	0.38^{b}	0.36^{a}	0.49^{d}	0.48^{d}	0.40^{c}
R_P	0.995	0.996	0.996	0.995	0.996	0.925	0.937	0.887	0.877	0.92

nvars: number of variables; *h*Mod: number of latent variables of the model; *RMSECV*: root mean square error of cross-validation; R_{CV}: correlation coefficient of cross-validation; *RMSEP*: root mean square error of prediction, R_P: correlation coefficient of prediction. Units: NIR – lignin (% w/w), UV – phorbol esters (mg g^{-1}), NMR – pentanol (%),GC – overall quality (not applied). Different superscript letters mean that the models are significantly different.

Table II.3. (continued).

	Vis/NIR - Xylene					XRF - Ni				
	Full	**OPS**	**GA**	***i*SPA**	**rPLS**	**Full**	**OPS**	**GA**	***i*SPA**	**rPLS**
nvars	316	45	59 – 63	175 – 285	27	261	30	79 – 86	79 – 105	248
*h*Mod			5					4		
RMSECV	0.21	0.14	0.06	0.19	0.15	0.173	0.149	0.117	0.246	0.172
R_{CV}	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
RMSEP	0.16^{c}	0.12^{a}	0.13^{b}	0.18^{d}	0.14^{b}	0.185^{b}	0.139^{a}	0.204^{c}	0.208^{c}	0.185^{b}
R_P	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
	QSAR - MMP-2 pIC_{50}					**MS - Peroxide Value**				
	Full	**OPS**	**GA**	***i*SPA**	**rPLS**	**Full**	**OPS**	**GA**	***i*SPA**	**rPLS**
nvars	439	195	48 – 117	373 – 439	26	106	40	26 – 53	69 – 91	104
*h*Mod			3					6		
RMSECV	0.45	0.4	0.31	0.45	0.33	0.44	0.39	0.28	0.48	0.49
R_{CV}	0.647	0.731	0.834	0.669	0.829	0.999	0.999	0.999	0.999	0.999
RMSEP	0.51^{c}	0.37^{a}	0.51^{c}	0.51^{c}	0.48^{b}	0.39^{b}	0.32^{a}	0.52^{c}	0.39^{b}	0.39^{b}
R_P	0.204	0.564	0.255	0.206	0.263	0.999	0.999	0.999	0.999	0.999

nvars: number of variables; *h*Mod: number of latent variables of the model; *RMSECV*: root mean square error of cross-validation; R_{CV}: correlation coefficient of cross-validation; *RMSEP*: root mean square error of prediction, R_P: correlation coefficient of prediction. Units: Vis/NIR – xylene (weight percent - %), XRF – Ni (%), QSAR – MMP-2 pIC_{50} (not applied), and MS – peroxide value (meq Kg^{-1}). MMP-2 pIC_{50}: Half-maximal inhibitory concentration negative logarithm of type 2 matrix metalloproteinases. Different superscript letters mean that the models are significantly different.

The *RMSEP* values are followed by superscript letters indicating the Tukey test results. Different letters mean that the models are significantly different. For MS (ethanol) and UV, all five PLS models (full and four variable selection models) differ significantly between then, being the OPS, the most predictive model. Only for fluorescence dataset, OPS and GA models were statistically equivalent. Furthermore, *RMSEP* values of full, OPS, GA, and rPLS do not differ significantly in Raman dataset.

In some cases, the least predictive model was not the full. For example, in voltammetry, NIR, GC, Vis/NIR, and XRF, *i*SPA models have the highest *RMSEP* value.

Also, the *RMSEP* was higher than full in rPLS for UV and GC datasets, and in GA for UV, GC, XRF, and MS (peroxide value).

In QSAR models, it is recommended to verify the possibility of chance correlation using the y-randomization test [36], where the y response was randomized five hundred times. Then, OPS and rPLS models do not show chance correlation. However, GA, *i*SPA, and full models presented chance correlation (results not shown).

Special attention should be given to MS (ethanol) dataset. Results for *i*SPA and rPLS models were not obtained using as the start point of selection the matrix with all two hundred variables because those algorithms were not able to perform the variable selection. In *i*SPA, the algorithm started the selection but paralyzed and did not finish the selection. For rPLS, the selection has not even begun; an error appears indicating that the **X** data, not present variance. From these, it can be presumed that both methods, *i*SPA, and rPLS, show the limitation to select variables in datasets like GC and MS, where the baseline does not have variance, and a peak suddenly arises. Thought, in this work, these algorithms perform the selection in a GC and another MS dataset. This can be explained by the presence of noise in the baseline, which confers variance to the entire set of variables. So, to perform the selection with *i*SPA and rPLS algorithms, the variables that do not present variance were eliminated and the full matrix now consists of 112 variables. This new **X** was used to build the full model and to start the variable selection.

In Figure II.8 is shown a bar plot with the *RMSEP* values and the number of selected variables of each variable selection algorithm for each dataset. These results are shown in percentage of increase or decrease regarding the full models. For OPS models, the percentage of decrease in *RMSEP* values ranged from 5 to 47% and the percentage of decrease in number of variables ranged from 56 to 99%. For GA and *i*SPA models, an error bar is shown because these two variable selection methods do not show stability when executing the selection more

than once. *i*SPA was reproducible only in MS (ethanol), UV, and NMR datasets. OPS and rPLS do not show any variance of selected variables in all twelve datasets, indicating that they are highly reproducible.

Figure II.9 presents the selected variables for all datasets. Each subplot shows the selected variables for each algorithm (OPS, GA, *i*SPA, and rPLS), and the frequency at which a variable was selected. The variable selection by OPS occurred, in general, where well-defined regions are observed, *i.e.*, peaks with rather clear physical meaning were selected. In some datasets, baseline regions are also selected, but despite this, the models are predictive. Besides that, the OPS selected few variables in all datasets, while the other methods selected almost all variables in some cases.

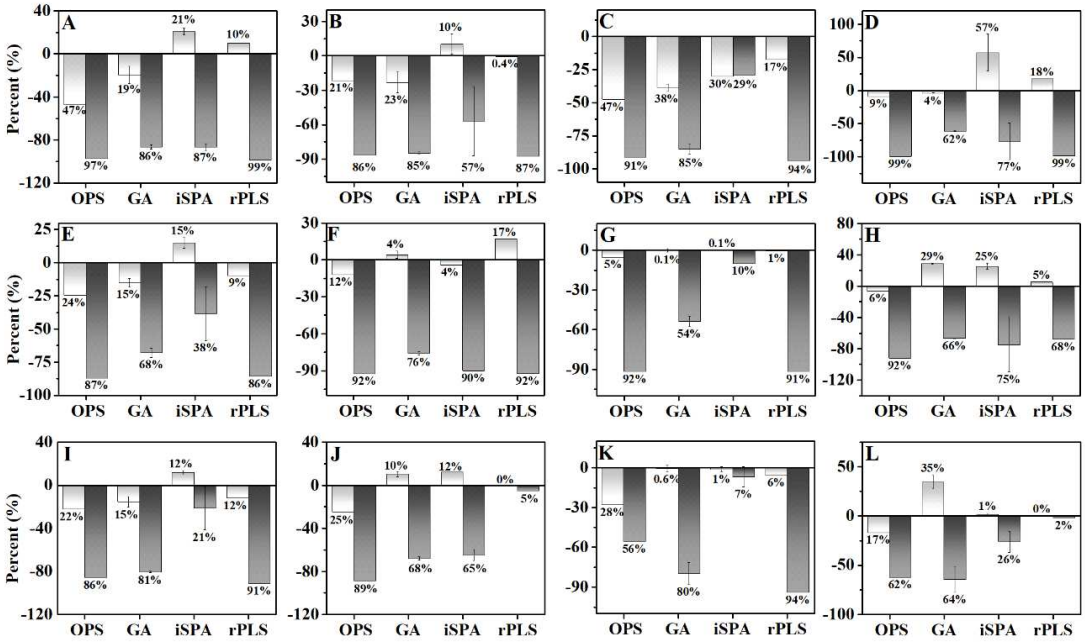

Figure II.8 *RMSEP* values of variable selection models (light gray bar) and the number of selected variables (dark gray bar). The values outside the bar are the percent of decrease (negative values) or increase (positive values) of *RMSEP* values or the number of variables regarding full model represented by the horizontal line on zero. (A) Voltammetry – ascorbic acid (μmol L^{-1}), (B) Fluorescence – catechol (μmol L^{-1}), (C) MS – ethanol (vol - %), (D) Raman – iodine value (g I_2 100g fat^{-1}), (E) NIR – lignin (% w/w), (F) UV – phorbol esters (mg g^{-1}), (G) NMR – pentanol (%), (H) GC – overall quality, (I) Vis/NIR – xylene (weight percent - %), (J) XRF – Ni (%), (K) QSAR – MMP-2 pIC_{50}, and (L) MS – peroxide value (meq Kg^{-1}). *RMSEP*: root mean square error of prediction.

The variables selected by the four methods do not match in most datasets. Few common variables are selected in all four methods (with the maximum frequency) as can be seen at the top of each subplot in Figure II.9. Therefore, the selection methods studied in this work do not converge to the same result of selected variables.

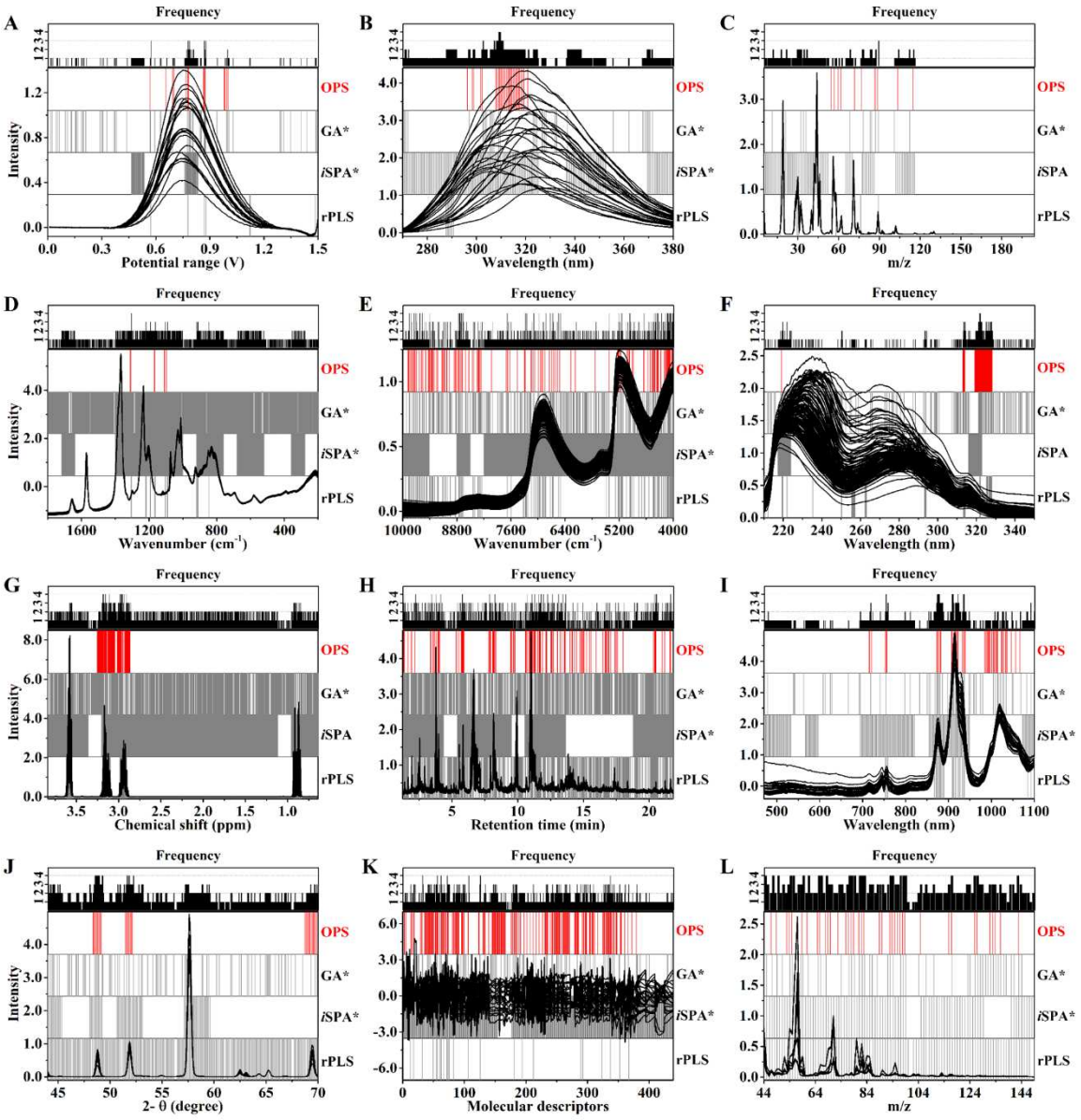

Figure II.9 (A) Voltammetry – ascorbic acid, (B) Fluorescence – catechol, (C) MS – ethanol, (D) Raman – iodine value, (E) NIR – lignin, (F) UV – phorbol esters, (G) NMR – pentanol, (H) GC – overall quality, (I) Vis/NIR – xylene, (J) XRF – Ni, (K) Autoscale QSAR – MMP-2 pIC50, and (L) MS – peroxide value with the variables selected (vertical lines) by OPS, GA, *i*SPA, and rPLS. The top graph in each subplot is the frequency that a variable was selected by the four variable selection methods. *Variables selected by the model with lowest *RMSEP*.

Even with some very similar selection strategies to OPS (rPLS uses the regression vector, and *i*SPA uses intervals), the other methods are not able to achieve the same result; they are very different. For example, in Figure II.9B, the variables selected by OPS are more informative than those selected by the other methods since these variables are in the region with higher intensity in the pure emission spectra of catechol (305 to 320 nm). In Figure II.9G, the RMN

spectra are composed by four signals. GA and *i*SPA selected variables throughout the spectra, including baseline. The rPLS and the OPS selected four and two signals, respectively.

The RMN signal between 3.00 to 3.25 chemical shift with high intensity is pentanol specific, and the OPS algorithm was able to select this specific signal, proving that OPS finds the region with higher selectivity. Thus, the new OPS was able to select more informative and predictive variables.

Conclusion

The new comprehensive approaches of OPS, *auto*OPS, *feed*OPS, and *i*OPS algorithms were developed and compared to the OPS*v*1 and the other three methods of variable selection concerning prediction capacity. The prediction performance of the new strategies outperformed the OPS*v*1. The new OPS algorithms were successfully applied to several types of dataset (sparse or non-sparse). They selected interpretative variables with greater predictability and were highly reproducible, selecting the same set of variables when executing the selection more than once. In general, the new strategies were able to reduce the number of variables in all datasets significantly. Besides that, they were better than the other methods in the external prediction of all datasets. Although the new approaches presented good results, there is not a best strategy nor informative vector suitable for all datasets or dataset type. A unique algorithm that automatically executes the three new OPS approaches is easily programmable. Overall, the OPS proved to be a universal and powerful method that significantly improved the prediction ability of the models, making it simpler to interpret, and showing excellent stability by selecting the same set of variables when executing the selection more than once.

References

[1] C.M. Andersen, R. Bro, Variable selection in regression-a tutorial, J. Chemom. 24 (2010) 728–737. doi:10.1002/cem.1360.

[2] B. Nadler, R.R. Coifman, The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration, J. Chemom. 19 (2005) 107–118. doi:10.1002/cem.915.

[3] R.G. Brereton, Introduction to multivariate calibration in analytical chemistry, Analyst. 125 (2000) 2125–2154. doi:10.1039/b003805i.

[4] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130. doi:10.1016/S0169-7439(01)00155-1.

[5] S. Wold, J. Trygg, A. Berglund, H. Antti, Some recent developments in PLS modeling, Chemom. Intell. Lab. Syst. 58 (2001) 131–150. doi:10.1016/S0169-7439(01)00156-3.

[6] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta. 667 (2010) 14–32. doi:10.1016/j.aca.2010.03.048.

[7] M. Goodarzi, Y. Vander Heyden, S. Funar-Timofei, Towards better understanding of feature-selection or reduction techniques for Quantitative Structure–Activity Relationship models, TrAC Trends Anal. Chem. 42 (2013) 49–63. doi:10.1016/j.trac.2012.09.008.

[8] A.A. Gomes, R.K.H. Galvão, M.C.U. Araújo, G. Véras, E.C. da Silva, The successive projections algorithm for interval selection in PLS, Microchem. J. 110 (2013) 202–208. doi:10.1016/j.microc.2013.03.015.

[9] R. Leardi, L. Nørgaard, Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, J. Chemom. 18 (2004) 486–497. doi:10.1002/cem.893.

[10] R. Leardi, A. Lupiáñez González, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemom. Intell. Lab. Syst. 41 (1998) 195–207. doi:10.1016/S0169-7439(98)00051-3.

[11] Å. Rinnan, M. Andersson, C. Ridder, S.B. Engelsen, Recursive weighted partial least squares (rPLS): an efficient variable selection method using PLS, J. Chemom. 28 (2014) 439–447. doi:10.1002/cem.2582.

[12] J.P.M. Andries, Y. Vander Heyden, L.M.C. Buydens, Predictive-Property-Ranked Variable Reduction with Final Complexity Adapted Models in Partial Least Squares Modeling for Multiple Responses, Anal. Chem. 85 (2013) 5444–5453. doi:10.1021/ac400339e.

[13] A.M.K. Pedro, M.M.C. Ferreira, Nondestructive Determination of Solids and Carotenoids in Tomato Products by Near-Infrared Spectroscopy and Multivariate Calibration, Anal. Chem. 77 (2005) 2505–2511. doi:10.1021/ac048651r.

[14] R.M. Balabin, S. V. Smirnov, Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data, Anal. Chim. Acta. 692 (2011) 63–72. doi:10.1016/j.aca.2011.03.006.

[15] I.R.N. de Oliveira, J. V. Roque, M.P. Maia, P.C. Stringheta, R.F. Teófilo, New strategy for determination of anthocyanins, polyphenols and antioxidant capacity of Brassica oleracea liquid extract using infrared spectroscopies and multivariate regression, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 194 (2018) 172–180. doi:10.1016/j.saa.2018.01.006.

[16] C. Assis, R.S. Ramos, L.A. Silva, V. Kist, M.H.P. Barbosa, R.F. Teófilo, Prediction of Lignin Content in Different Parts of Sugarcane Using Near-Infrared Spectroscopy (NIR), Ordered Predictors Selection (OPS), and Partial Least Squares (PLS), Appl. Spectrosc. 71 (2017) 2001–2012. doi:10.1177/0003702817704147.

[17] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression, J. Chemom. 23 (2009) 32–48. doi:10.1002/cem.1192.

[18] J. Yuan, S. Yu, T. Zhang, X. Yuan, Y. Cao, X. Yu, X. Yang, W. Yao, QSPR models for predicting generator-column-derived octanol/water and octanol/air partition coefficients of polychlorinated biphenyls, Ecotoxicol. Environ. Saf. 128 (2016) 171–180. doi:10.1016/j.ecoenv.2016.02.022.

[19] C.S.W. Miaw, C. Assis, A.R.C.S. Silva, M.L. Cunha, M.M. Sena, S.V.C. de Souza, Determination of main fruits in adulterated nectars by ATR-FTIR spectroscopy combined with multivariate calibration and variable selection methods, Food Chem. 254 (2018) 272–280. doi:10.1016/j.foodchem.2018.02.015.

[20] C. Assis, L.S. Oliveira, M.M. Sena, Variable Selection Applied to the Development of a Robust Method for the Quantification of Coffee Blends Using Mid Infrared Spectroscopy, Food Anal. Methods. 11 (2018) 578–588. doi:10.1007/s12161-017-1027-

7.
[21] Í.P. Caliari, M.H.P. Barbosa, S.O. Ferreira, R.F. Teófilo, Estimation of cellulose crystallinity of sugarcane biomass using near infrared spectroscopy and multivariate analysis methods, Carbohydr. Polym. 158 (2017) 20–28. doi:10.1016/j.carbpol.2016.12.005.
[22] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta. 185 (1986) 1–17. doi:10.1016/0003-2670(86)80028-9.
[23] R.W. Kennard, L.A. Stone, Computer Aided Design of Experiments, Technometrics. 11 (1969) 137–148. doi:10.1080/00401706.1969.10490666.
[24] R.K.H. Galvão, M.C.U. Araújo, W.D. Fragoso, E.C. Silva, G.E. José, S.F.C. Soares, H.M. Paiva, A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm, Chemom. Intell. Lab. Syst. 92 (2008) 83–91. doi:10.1016/j.chemolab.2007.12.004.
[25] O.O. Soyemi, M.A. Busch, K.W. Busch, Multivariate Analysis of Near-Infrared Spectra Using the G-Programming Language, J. Chem. Inf. Comput. Sci. 40 (2000) 1093–1100. doi:10.1021/ci000447r.
[26] N. (Klaas) M. Faber, R. Bro, Standard error of prediction for multiway PLS, Chemom. Intell. Lab. Syst. 61 (2002) 133–149. doi:10.1016/S0169-7439(01)00204-0.
[27] M. Dyrby, S.B. Engelsen, L. Nørgaard, M. Bruhn, L. Lundsberg-Nielsen, Chemometric Quantitation of the Active Substance (Containing C≡N) in a Pharmaceutical Tablet Using Near-Infrared (NIR) Transmittance and NIR FT-Raman Spectra, Appl. Spectrosc. 56 (2002) 579–585. doi:10.1366/0003702021955358.
[28] R. Kiralj, M.M.C. Ferreira, A priori molecular descriptors in QSAR: a case of HIV-1 protease inhibitors, J. Mol. Graph. Model. 21 (2003) 435–448. doi:10.1016/S1093-3263(02)00201-2.
[29] R.F. Teófilo, Métodos quimiométricos em estudos eletroquímicos de fenóis sobre filmes de diamante dopado com boro, Thesis (Doutorado em Química) - Universidade Estadual de Campinas, 2007 (Brazil).
[30] T. Skov, D. Ballabio, R. Bro, Multiblock variance partitioning: A new approach for comparing variation in multiple data blocks, Anal. Chim. Acta. 615 (2008) 18–29. doi:10.1016/j.aca.2008.03.045.
[31] L.B. Lyndgaard, K.M. Sørensen, F. Berg, S.B. Engelsen, Depth profiling of porcine adipose tissue by Raman spectroscopy, J. Raman Spectrosc. 43 (2012) 482–489. doi:10.1002/jrs.3067.
[32] J. V Roque, L.A.S. Dias, R.F. Teófilo, Multivariate Calibration to Determine Phorbol Esters in Seeds of Jatropha curcas L. Using Near Infrared and Ultraviolet Spectroscopies, J. Braz. Chem. Soc. 28 (2017) 1506–1516. doi:10.21577/0103-5053.20160332.
[33] H. Winning, F.H. Larsen, R. Bro, S.B. Engelsen, Quantitative analysis of NMR spectra with chemometrics, J. Magn. Reson. 190 (2008) 26–32. doi:10.1016/j.jmr.2007.10.005.
[34] J.S. Ribeiro, F. Augusto, T.J.G. Salva, M.M.C. Ferreira, Prediction models for Arabica coffee beverage quality based on aroma analyses and chemometrics, Talanta. 101 (2012) 253–260. doi:10.1016/j.talanta.2012.09.022.
[35] Y. Wang, X. Zhao, B.R. Kowalski, X-Ray Fluorescence Calibration with Partial Least-Squares, Appl. Spectrosc. 44 (1990) 998–1002. doi:10.1366/0003702904086867.
[36] E.B. de Melo, A QSAR Study of Matrix Metalloproteinases Type 2 (MMP-2) Inhibitors with Cinnamoyl Pyrrolidine Derivatives, Sci. Pharm. 80 (2012) 265–281. doi:10.3797/scipharm.1112-21.

Chapter III

Computational performance of variable selection methods

Abstract

Simulated datasets were used to evaluate the computational performance of the ordered predictors selection (OPS) and other variable selection methods, *i.e.*, genetic algorithm (GA), the interval successive projections algorithm for PLS (*i*SPA), and recursive weighted partial least squares (rPLS). The OPS methods studied were the automatic OPS, the feedback OPS, and the OPS intervals. All methods were evaluated by using a central composite design varying the matrix dimensions and the number of latent variables. The computational performance of the OPS methods was mainly influenced by the number columns, as well as the GA. On the other hand, *i*SPA and rPLS were mainly influenced by the number of rows. In general, the performance times of the OPS methods were similar to the other methods. Among the six methods analyzed, rPLS was the fastest, followed by GA, OPS approaches, and *i*SPA. However, even though the OPS methods were not the fastest ones, in the literature, they better improved the models' quality of several real datasets in comparison to other variable selection methods.

Keywords: Variable Selection, Running Time, Simulated Datasets, Cross-Validation, PLS.

Introduction

Multivariate calibration has been extensively used in several science fields. This type of method takes into consideration several variables and a property of interest to build a quantitative relationship. There are several multivariate calibration methods, such as multiple linear regression (MLR), support vector machine regression (SVMR) and partial least squares regression (PLS), the latter being the most used in chemometrics [1]. The considerable interest in PLS is mainly due to its capability to analyze strongly collinear, noisy, and numerous variables and the possibility to simultaneously model several properties [2].

Nowadays, most datasets used in chemometric approaches are enormous, having hundreds or thousands of variables which are highly correlated. Although PLS works well for these datasets, it may include irrelevant, nonlinear, and noisy variables in the model, jeopardizing its interpretation. This evidence suggests that data reduction, such as variable selection, is necessary to improve the model [3].

Variable selection is a key step in multivariate calibration. It reduces data dimensionality, minimizing the influence of irrelevant variables, improving the model´s quality, and simplifying its interpretation. Several results in literature [4–7] recognizes the critical role played by variables selection methods [8–13]. However, some methods may spend a long time to select variables. Then, the computational time must be taken into account when choosing the variable selection method to be used, mostly when dealing with large datasets.

Previous research has established that the computational time and performance of different PLS algorithms differ according to their mathematical approaches to find the PLS regression vector [14–16]. Besides, the PLS calculation time is also affected by the dataset size [14,16]. Thus, there is some evidence that suggests that a large set of variables may also affect the computational time and performance of the variable selection step.

A fast variable selection algorithm that provides interpretative and predictive variables is highly demanded. Over the years, several variable selection methods were developed to achieve these requirements. Among these methods is the ordered predictors selection (OPS) algorithm, initially presented by Teófilo et al. [12] and recently updated by Roque et al. [13]. The OPS methods are simple and efficient for the selection of any variable type [17–19], providing a high ability to improve the prediction of multivariate calibration models with few variables [20–22].

The purpose of this work was to evaluate the computational running time of several feature selection algorithms using simulated datasets of different sizes. The OPS methods [13]

were compared with other three variable selection methods available in the literature, *i.e.*, genetic algorithm (GA) [10], intervals successive projections algorithm for PLS (*i*SPA) [23], and recursive weighted partial least squares (rPLS) [11].

Variable selection methods

Ordered Predictors Selection (OPS)

The core of OPS sorts the variables based on informative vectors and systematically investigates multivariate calibration models to find the most relevant set of interpretable variables by comparing the cross-validation parameters of the models [12,13]. First, an informative vector containing information about the best independent variables for prediction is obtained. Then, the independent variables of the **X** matrix are differentiated according to their corresponding absolute values in the informative vector. An initial subset of variables (window) is defined to build and evaluate the first model. After, this initial subset is extended by the addition of a fixed number of variables (increment) over the window, and a new model is built and evaluated. Further increments are added until all, or a percentage of the variables are considered. Cross-validation parameters, such as the root mean square error of cross-validation (*RMSECV*) and the correlation coefficient of cross-validation (R_{cv}), are calculated for each model. Finally, the best variable subset is defined comparing the quality parameters calculated during the validation step. Figure III.1 shows a flowchart for the core of OPS.

Three new strategies to make a more versatile OPS were recently presented [13]. The first one is the automatic OPS (*auto*OPS) that automatically performs all steps of core OPS using several informative vectors providing the best results. The second is the feedback OPS (*feed*OPS) [13], wherein pre-selected variables return to a new selection run until achieving a convergence criterion. This method runs the *auto*OPS in loop constrained by a rule. The last strategy is the interval OPS (*i*OPS), in which *auto*OPS or *feed*OPS is applied in subdivisions of the full array. These approaches are presented in Figure III.2.

Genetic Algorithm (GA)

The GA is a method used for solving optimization problems based on natural selection and genetic processes that mimic biological evolution [10]. Figure III.3A shows a scheme of the GA steps.

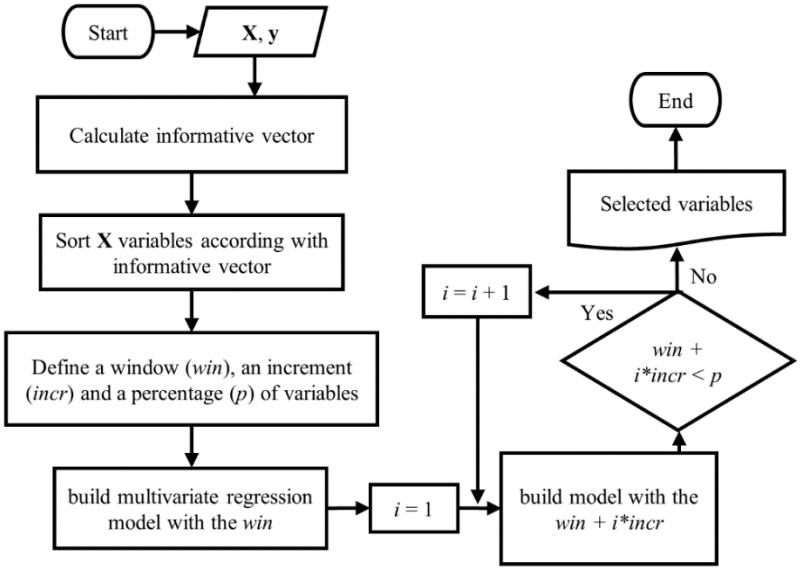

Figure III.1. Algorithm flowchart of the core of OPS method.

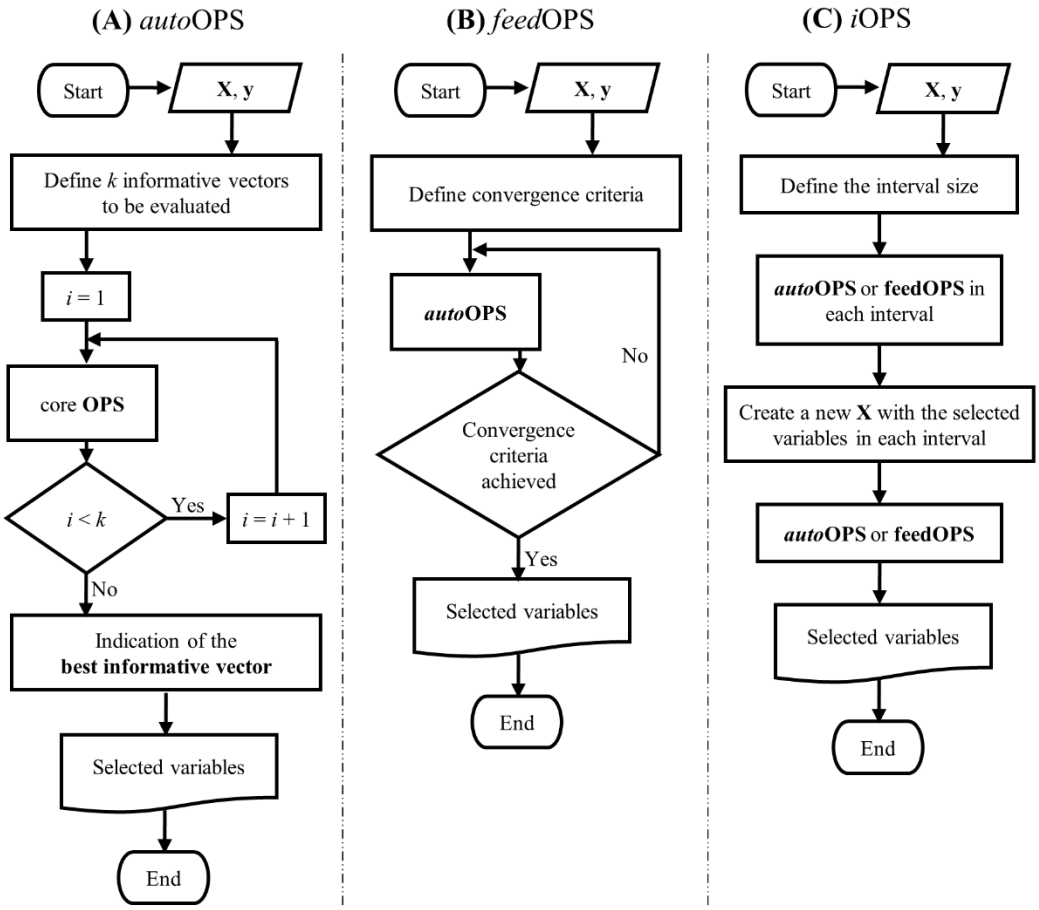

Figure III.2. Algorithm flowchart of (A) *auto*OPS, (B) *feed*OPS, and (C) *i*OPS methods.

Initially, a starting population is constituted for a series of different individuals, and the number of individuals is defined as the population with a defined size. These individuals are reproduced, crossed, mutated, and evaluated according to the parameters introduced into the algorithm until achieve some termination criterion. This criterion can be based on a lack of improvement of the model (*e.g. RMSECV*) or simply on a maximum number of generations (*i*). At the end of the process, a vector consisting of zeros and ones indicates whether variables should be included (1) or not (0).

Intervals Successive Projections Algorithm (iSPA)

The *i*SPA combines the noise-reduction properties of PLS with the possibility of discarding non-informative variables with SPA-MLR algorithm [24]. In this method, it is assumed that the variables have been divided into non-overlapping intervals (*i*), usually, with the same length. The number of intervals is one of the settings inputs. The columns of **X** are partitioned according to the intervals of variables previously defined. A single representative variable of each interval (w_i) is used in a sequence of projection operations that result in the creation of chains. This process is repeated until all representative variables have been projected in each orthogonal space of w_i. PLS models are built using leave-one-out cross-validation for each combination of intervals, and the best combination of intervals is then chosen by the smallest *RMSECV* [23]. This process is illustrated in Figure III.3B.

Recursive Weighted Partial Least Squares (rPLS)

The rPLS method uses the dependent variable **y** to guide the variable weighting recursively (Figure III.3C). This method iteratively reweights the variables using the regression vector ($\boldsymbol{b}_i$) calculated by PLS [11]. This process is repeated until *RMSECV* converges to a minimum value. The variables are eliminated based on a filter parameter, which variables with a large difference between the smallest and the largest regression coefficient are removed. The number of iterations and the filter parameter are part of the settings input.

Experimental

This section is about the evaluation of computational performance of six variable selection methods algorithms: *auto*OPS [13], *feed*OPS [13], *i*OPS [13], GA [10], *i*SPA [23], and rPLS [11]. The evaluation was made through an experimental design using simulated datasets. All calculations were performed in MATLAB 2019a (Math Works, Natick, USA) environment.

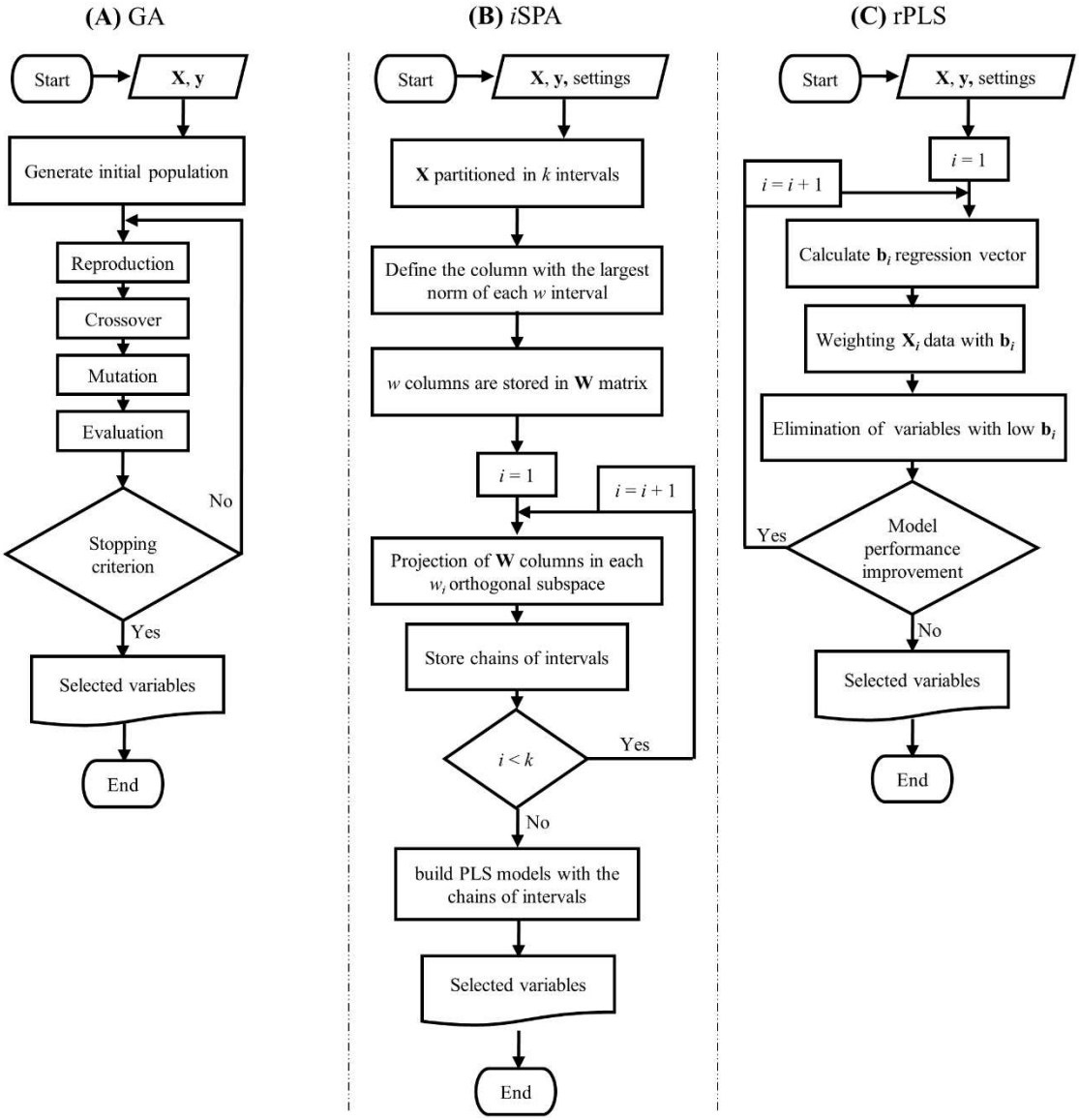

Figure III.3. Algorithm flowchart of (A) GA, (B) *i*SPA, and (C) rPLS methods.

Simulated Datasets

Five-level central composite design (CCD) [25] was carried out. The PLS regression was used to build multivariate models for all variable selection methods. Variable selection was performed using independent (**X**) and dependent (**y**) variables generated using a pseudo-random function of MATLAB 2019a. The number of rows (samples) of **X** and **y**, number of columns (variables) of **X,** and the number of latent variables (*nlvs*) of the PLS models were the factors evaluated in this design. Table III.1 shows the factors and levels investigated.

Table III.1. Factors with coded and decoded levels for five-level CCD designs.

	Levels				
Factors	**-1.68**	**-1**	**0**	**1**	**1.68**
Rows (Samples)	109	310	605	900	1101
Columns (Variables)	227	5000	12000	19000	23773
nlvs	2	6	12	17	21

nlvs: number of latent variables.

The response of CCD was the running time of the variable selection algorithms. This response was acquired using functions of MATLAB (*tic* and *toc*) that measure the elapsed time. All data were processed by a dedicated computer Intel Core i5, with HD SATA 500 GB and RAM 6 GB with the Microsoft Windows 8.1 operational system.

Therefore, a total of six CCD designs were performed. Three of them refer to new OPS approaches (*auto*OPS, *feed*OPS, and *i*OPS), and the other three refer to GA, *i*SPA, and rPLS algorithms. Each CCD design consisted of nineteen experiments, including eight factorial experimental points, six axial experimental points, and five replicates in the experimental central point.

Statistical evaluation by mean square residual error was performed considering regression coefficients calculated by the following equation:

$$\mathbf{b} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \qquad \textbf{(III.1)}$$

where **b** is the regression coefficients, **X** is the matrix of coded factors, and **y** is the vector of design response. All calculations of CCD design were performed using spreadsheets, according to Teófilo and Ferreira [25].

Algorithms Parameters

The default settings of each variable selection algorithm were used to evaluate the computational performance. The inputs consisted only of **X**, **y**, and the number of latent variables (*nlvs*) for all algorithms.

The OPS algorithms were applied using the informative vectors of the 'main' option, window of 10 and increments of 5 variables, 100% of variables were tested, and venetian blinds cross-validation was applied with splits of 10. In *feed*OPS, a minimum difference of 2% between two consecutive *RMSECV*s and a maximum of ten loops were used as convergence criteria. In *i*OPS, the full matrix was divided into intervals of 50 variables, and the option to run the selection using *auto*OPS in each interval was used. Additionally, *h*OPS, the number of latent

variables used to calculate informative vectors (more information can be found elsewhere [13]), was calculated for each interval in *i*OPS. The OPS variable selection was performed using the algorithms for MATLAB available at www.deq.ufv.br/chemometrics.

The GA was applied using the following conditions: a population of 64, a maximum generation of 100, a mutation rate of 0.005, a window width of 1, convergence of 50, 30 terms included at initiation, cross-over rule of 2, contiguous block cross-validation was applied with the number of subsets to divide data into for cross-validation of 5, cross-validation iteration of 1, and replicate runs to perform of 1. The GA variable selection was performed using the *.m* functions from PLS-Toolbox 8.2 (Eigenvector Research, Inc. Wenatchee, USA).

The *i*SPA algorithm performs leave-one-out cross-validation. By default settings, the full **X** matrix was divided into ten intervals, being ten the maximum number of intervals to be selected. The authors provided this algorithm for MATLAB.

The rPLS method was applied with random cross-validation with splits of 10, infinite iterations, and the large difference between the smallest and the largest regression coefficient allowed was set as 1000. This algorithm for MATLAB was downloaded from www.models.life.ku.dk/algorithms.

Results and Discussion

Figure III.4 shows the results of running time, in minutes, for the six CCD designs. From Figure III.4A, a short time is observed for rPLS (blue circles) in all nineteen experimental points. GA algorithm (green squares) also presented a relatively short time when compared to higher times showed by OPS (red, blue and orange triangles) and *i*SPA (pink diamond) algorithms. Figure III.4B shows that the running time behavior is similar for all algorithms in most of CCD design assays. The time was normalized by using the Euclidean norm to facilitate the visualization and comparison between algorithms.

Table III.2 summarizes the CCD running time results obtained for *auto*OPS, *feed*OPS, *i*OPS, GA, *i*SPA, and rPLS algorithms. The minimum and maximum values of running time obtained for each CCD design are highlighted in bold.

The rPLS algorithm spent a short time in all CCD experiments, between 0.19 and 4.13 minutes. The three OPS algorithms showed minimum running times between 0.11 and 0.24 minutes in datasets with 605 rows and 227 columns, while GA and *i*SPA took more than two minutes.

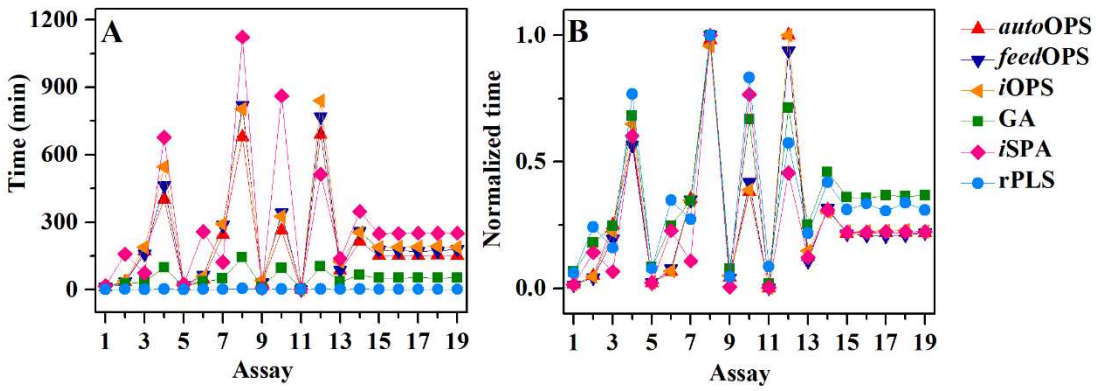

Figure III.4. Comparison of running times of *auto*OPS, *feed*OPS, *i*OPS, GA, *i*SPA, and rPLS algorithms in minutes (A) and in normalized time by Euclidean norm (B).

*Auto*OPS and *i*OPS algorithms spent 690 and 839 minutes (ten and fourteen hours), respectively, in the experimental point of CCD design with a max number of columns in **X** (assay 12). *Feed*OPS spent more running time (819 minutes) in the experimental with 900 rows and 19000 columns. In the same experimental point, GA spent 143 minutes or about two and a half hours. However, GA is an algorithm that needs to be optimized to yield good results in variable selection. In this computational performance test, GA was not optimized, spending much less time than OPS algorithms. The *i*SPA algorithm spent 1124 minutes (about nineteen hours) to perform selection in a large matrix (both rows and columns in +1 level). The long time is probably related to the leave-one-out cross-validation performed by this method. Then, the larger the matrix size, both rows and columns, the longer the running time of selection.

Statistical analysis by mean square residual error and the significance level (α) at 0.05 indicates that all linear regression coefficients are positively significant for all algorithms, as shown in the Pareto chart (Figure III.5). Then, increasing the number of columns, rows, and *nlvs* increases the running time of the three new OPS approaches, GA, *i*SPA, and rPLS. In Figure III.5A, III.5B, and III.5C, the Pareto charts for *auto*OPS, *feed*OPS, and *i*OPS algorithms, respectively, are presented. In Figure III.5D, III.5E, and III.5F, the Pareto charts for GA, *i*SPA, and rPLS, respectively, are shown.

For all three OPS algorithms and GA, the number of columns showed a higher effect, followed by the number of rows. This was expected because variable selection methods deal with the number of columns and spend more time performing selection when the matrix has more columns. However, for *i*SPA and rPLS, the opposite occurred. The number of rows showed a higher effect, followed by the number of columns.

Table III.2. Five-level CCD and time results in minutes for variable selection methods.

Assay	Variables and levels			Time (min)					
	Rows	**Columns**	***nlvs***	***auto*OPS**	***feed*OPS**	***i*OPS**	**GA**	***i*SPA**	**rPLS**
1	310 (-1)	5000 (-1)	6 (-1)	10.63	10.64	14.12	9.41	15.87	0.25
2	900 (+1)	5000 (-1)	6 (-1)	33.07	34.11	38.54	25.61	158.42	1.00
3	310 (-1)	19000 (+1)	6 (-1)	172.18	159.20	188.25	35.27	73.73	0.66
4	900 (+1)	19000 (+1)	6 (-1)	401.57	463.38	546.39	97.76	677.88	3.17
5	310 (-1)	5000 (-1)	17 (+1)	14.35	18.60	19.13	11.88	23.72	0.32
6	900 (+1)	5000 (-1)	17 (+1)	44.34	64.05	57.02	35.46	256.80	1.44
7	310 (-1)	19000 (+1)	17 (+1)	243.83	285.71	292.31	49.58	121.41	1.12
8	900 (+1)	19000 (+1)	17 (+1)	677.61	**819.48**	804.38	**143.38**	**1124.20**	**4.13**
9	109 (-1.68)	12000 (0)	12 (0)	29.77	32.36	42.04	11.09	5.69	**0.19**
10	1101 (+1.68)	12000 (0)	12 (0)	263.04	342.54	326.82	95.62	861.95	3.45
11	605 (0)	227 (-1.68)	12 (0)	**0.14**	**0.11**	**0.24**	**2.50**	**3.40**	0.35
12	605 (0)	23773 (+1.68)	12 (0)	**690.64**	768.96	**839.83**	102.26	512.68	2.38
13	605 (0)	12000 (0)	2 (-1.68)	86.22	87.27	124.56	35.97	137.33	0.90
14	605 (0)	12000 (0)	21 (+1.68)	213.79	259.77	254.64	65.90	347.70	1.74
15	605 (0)	12000 (0)	12 (0)	149.62	175.98	188.62	51.51	248.93	1.28
16	605 (0)	12000 (0)	12 (0)	150.96	172.32	189.46	51.01	249.66	1.38
17	605 (0)	12000 (0)	12 (0)	150.06	170.13	190.41	52.51	251.45	1.26
18	605 (0)	12000 (0)	12 (0)	151.93	172.52	192.22	52.07	250.83	1.40
19	605 (0)	12000 (0)	12 (0)	150.06	179.07	187.88	52.53	250.95	1.28

Values in parenthesis are decoded levels for statistical analysis. Bold values are the minimum and the maximum of each algorithm. *nlvs*: number of latent variables.

From this, it can be assumed that *i*SPA and rPLS running times are significantly influenced by cross-validation since this process is directly related to the number of matrix rows. The *i*SPA algorithm performs the leave-one-out cross-validation, that is a slow method, and this may influence the number of rows having a more significant effect than the number of columns in the *i*SPA algorithm.

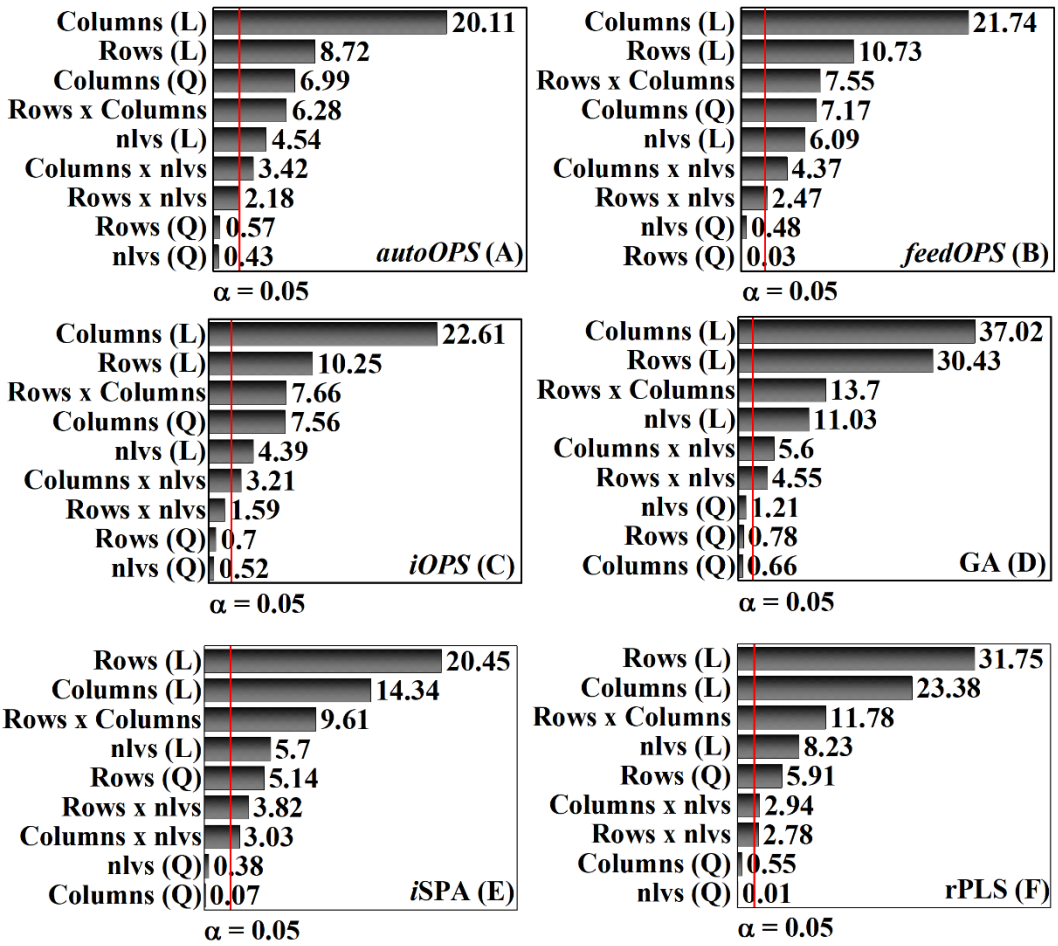

Figure III. 5. Pareto chart for the standardized main effects in the central composite design of *auto*OPS (A), *feed*OPS (B), *i*OPS (C), GA (D), *i*SPA (E), and rPLS (F). Red line is the significance level (α) at 0.05. Linear (L) and quadratic (Q) coefficients. Number of latent variables (*nlvs*).

In all OPS approaches, the significant quadratic coefficient for the number of columns indicates that the running time increases quadratically with the increase of columns. Moreover, this coefficient was significant only for OPS, indicating that the new algorithms are more affected by the increase of columns than the other methods tested.

A profile report for each variable selection method was generated by *profile* MATLAB function using performance time. The random **X** and **y** data created in the first CCD design

experimental point (310 rows, 5000 columns) and six latent variables were used. These reports were used to evaluate the time spent in subfunctions used in each variable selection method. Each algorithm performs the search for more predictive and informative variables using different strategies based on how they combine the variable selection and the model building. Although all of them use the cross-validation procedure as internal validation and build PLS models, the difference, however, lays on the different cross-validation strategies and PLS algorithms employed, as can be seen in Table III.3.

In GA and OPS algorithms, the search for variables takes longer than the execution of cross-validation (CV) and PLS, being these methods more affected by the number of columns or variables in **X** data, consistent with the Pareto results. However, rPLS and *i*SPA spent more time in PLS and CV procedures, respectively. About 64% of the time spent in *i*SPA algorithm refers to leave-one-out cross-validation, explaining why the rows numbers of **X** data has more influence in running time. In rPLS, 74% of the time was spent in PLS calculation, being consistent with the fact that this algorithm works in a loop to reweight the regression vector building several PLS models.

Table III.3. Profile report for OPS, GA, *i*SPA and rPLS with percent of time spent of cross-validation and PLS algorithms.

	Variable selection	**CV**		**PLS**		
	time (%)	type	splits	time (%)	type	time (%)
***auto*OPS**	40.06	venetian blinds	10	21.96	BIDIAG	37.98
***feed*OPS**	39.84	venetian blinds	10	21.91	BIDIAG	38.25
***i*OPS**	45.88	venetian blinds	10	21.57	BIDIAG	32.55
GA	65.72	contiguos block	5	9.52	SIMPLS	24.76
***i*SPA**	2.97	leave-one-out	310	64.59	NIPALS	32.44
rPLS	18.72	Random	10	7.28	BIDIAG	74.00

CV: cross-validation; BIDIAG: bidiagonalization algorithm; NIPALS: nonlinear iterative partial least squares.

OPS and rPLS algorithms use the bidiagonalization algorithm (BIDIAG) proposed by Manne [26], GA uses the SIMPLS algorithm developed by De Jong [27], and *i*SPA employs nonlinear iterative partial least squares (NIPALS) [28]. Martins et al. [14] evaluated the computational time of several PLS algorithms, and the BIDIAG algorithm was considered faster than SIMPLS, which was faster than NIPALS. The higher time performance of NIPALS

mentioned in literature also corroborates with the high running time spent by *i*SPA algorithm once 32% of its time is spent in PLS calculations.

Variable selection algorithms can be classified into three different classes, according to Mehmood et al. [3]. As OPS algorithms are an iterative procedure based on a greedy exploration that searches for an optimal subset of variables for each latent variable of the PLS model, they are classified as an embedded method. These embedded methods usually expend more computational time since the variable selection is an integrated step of modeling. Then, OPS algorithms spent considerable time to select variables as they perform a longing search for the best variables. GA is classified as a randomized wrapper method, which uses a threshold to select variables in an iterative way, where the model building is wrapped within the variable selection step. Wrapper methods tend to be computationally demanding. *i*SPA and rPLS algorithms are classified as filter methods, which are based only on a threshold to select variables. These methods are particularly useful in computation time, but they tend to select redundant variables. Besides, a recently published work [13] confirmed that the OPS method selected more interpretative and predictive variables, while rPLS was the fastest, but not the most predictive in most cases. Besides that, *i*SPA was the slowest algorithm and also was not predictive, selecting non-interpretative and predictive variables [13].

In addition, the results presented in this work could be used to improve the performance of variable selection algorithms and assist in the choice of the best pathway to develop fast algorithms.

Conclusion

The choice of the variable selection method for multivariate regression is an important step to be considered due to the significant differences in running time of algorithms presented in this work. The matrix dimensions have a more pronounced effect in running time, while the number of latent variables was the minor effect. The OPS algorithms were able to perform the variable selection in small-size matrices in short running times, and they were affected mainly by the number of columns, spending more time performing selection on data sets with a high number of variables. GA was mainly affected by the number of columns; however, *i*SPA and rPLS algorithms were predominantly affected by the number of rows. Concerning the average running times, the rPLS was the fastest, followed by GA, *auto*OPS, *feed*OPS, *i*OPS, and *i*SPA algorithms. Although the OPS approaches were not the fastest methods, they improved the

prediction and interpretative ability of several real datasets in the literature in comparison to other variable selection methods.

References

[1] H. Martens, T. Naes, T. Naes, Multivariate calibration, John Wiley & Sons, 1992.
[2] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130. doi:10.1016/S0169-7439(01)00155-1.
[3] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in Partial Least Squares Regression, Chemom. Intell. Lab. Syst. 118 (2012) 62–69. doi:10.1016/j.chemolab.2012.07.010.
[4] R.M. Balabin, S. V. Smirnov, Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data, Anal. Chim. Acta. 692 (2011) 63–72. doi:10.1016/j.aca.2011.03.006.
[5] J.P.M. Andries, Y. Vander Heyden, L.M.C. Buydens, Predictive-Property-Ranked Variable Reduction with Final Complexity Adapted Models in Partial Least Squares Modeling for Multiple Responses, Anal. Chem. 85 (2013) 5444–5453. doi:10.1021/ac400339e.
[6] A.M.K. Pedro, M.M.C. Ferreira, Nondestructive Determination of Solids and Carotenoids in Tomato Products by Near-Infrared Spectroscopy and Multivariate Calibration, Anal. Chem. 77 (2005) 2505–2511. doi:10.1021/ac048651r.
[7] C. Assis, R.S. Ramos, L.A. Silva, V. Kist, M.H.P. Barbosa, R.F. Teófilo, Prediction of Lignin Content in Different Parts of Sugarcane Using Near-Infrared Spectroscopy (NIR), Ordered Predictors Selection (OPS), and Partial Least Squares (PLS), Appl. Spectrosc. 71 (2017) 2001–2012. doi:10.1177/0003702817704147.
[8] M.C.U. Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, Chemom. Intell. Lab. Syst. 57 (2001) 65–73. doi:10.1016/S0169-7439(01)00119-8.
[9] R. Leardi, L. Nørgaard, Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, J. Chemom. 18 (2004) 486–497. doi:10.1002/cem.893.
[10] R. Leardi, A. Lupiáñez González, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemom. Intell. Lab. Syst. 41 (1998) 195–207. doi:10.1016/S0169-7439(98)00051-3.
[11] Å. Rinnan, M. Andersson, C. Ridder, S.B. Engelsen, Recursive weighted partial least squares (rPLS): an efficient variable selection method using PLS, J. Chemom. 28 (2014) 439–447. doi:10.1002/cem.2582.
[12] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression, J. Chemom. 23 (2009) 32–48. doi:10.1002/cem.1192.
[13] J. V. Roque, W. Cardoso, L.A. Peternelli, R.F. Teofilo, R.F. Teófilo, Comprehensive new approaches for variable selection using ordered predictors selection, Anal. Chim. Acta. (2019). doi:10.1016/j.aca.2019.05.039.
[14] J.P.A. Martins, R.F. Teófilo, M.M.C. Ferreira, Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets, J. Chemom. (2010) n/a-n/a. doi:10.1002/cem.1309.
[15] W. Wu, R. Manne, Fast regression methods in a Lanczos (or PLS-1) basis. Theory and

applications, Chemom. Intell. Lab. Syst. 51 (2000) 145–161. doi:10.1016/S0169-7439(00)00063-0.

[16] M. Andersson, A comparison of nine PLS1 algorithms, J. Chemom. 23 (2009) 518–529. doi:10.1002/cem.1248.

[17] J. Yuan, S. Yu, T. Zhang, X. Yuan, Y. Cao, X. Yu, X. Yang, W. Yao, QSPR models for predicting generator-column-derived octanol/water and octanol/air partition coefficients of polychlorinated biphenyls, Ecotoxicol. Environ. Saf. 128 (2016) 171–180. doi:10.1016/j.ecoenv.2016.02.022.

[18] C.S.W. Miaw, C. Assis, A.R.C.S. Silva, M.L. Cunha, M.M. Sena, S.V.C. de Souza, Determination of main fruits in adulterated nectars by ATR-FTIR spectroscopy combined with multivariate calibration and variable selection methods, Food Chem. 254 (2018) 272–280. doi:10.1016/j.foodchem.2018.02.015.

[19] M. Goodarzi, Y. Vander Heyden, S. Funar-Timofei, Towards better understanding of feature-selection or reduction techniques for Quantitative Structure–Activity Relationship models, TrAC Trends Anal. Chem. 42 (2013) 49–63. doi:10.1016/j.trac.2012.09.008.

[20] C. Assis, L.S. Oliveira, M.M. Sena, Variable Selection Applied to the Development of a Robust Method for the Quantification of Coffee Blends Using Mid Infrared Spectroscopy, Food Anal. Methods. 11 (2018) 578–588. doi:10.1007/s12161-017-1027-7.

[21] Í.P. Caliari, M.H.P. Barbosa, S.O. Ferreira, R.F. Teófilo, Estimation of cellulose crystallinity of sugarcane biomass using near infrared spectroscopy and multivariate analysis methods, Carbohydr. Polym. 158 (2017) 20–28. doi:10.1016/j.carbpol.2016.12.005.

[22] I.R.N. de Oliveira, J. V. Roque, M.P. Maia, P.C. Stringheta, R.F. Teófilo, New strategy for determination of anthocyanins, polyphenols and antioxidant capacity of Brassica oleracea liquid extract using infrared spectroscopies and multivariate regression, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 194 (2018) 172–180. doi:10.1016/j.saa.2018.01.006.

[23] A.A. Gomes, R.K.H. Galvão, M.C.U. Araújo, G. Véras, E.C. da Silva, The successive projections algorithm for interval selection in PLS, Microchem. J. 110 (2013) 202–208. doi:10.1016/j.microc.2013.03.015.

[24] S.F.C. Soares, A.A. Gomes, M.C.U. Araujo, A.R.G. Filho, R.K.H. Galvão, The successive projections algorithm, TrAC Trends Anal. Chem. 42 (2013) 84–98. doi:https://doi.org/10.1016/j.trac.2012.09.006.

[25] R.F. Teófilo, M.M.C. Ferreira, Quimiometria II: Planilhas eletrônicas para cálculus de planejamentos experimentais, um tutorial, Quim. Nova. 29 (2006) 338–350. doi:10.1590/S0100-40422006000200026.

[26] R. Manne, Analysis of two partial-least-squares algorithms for multivariate calibration, Chemom. Intell. Lab. Syst. 2 (1987) 187–197. doi:10.1016/0169-7439(87)80096-5.

[27] S. de Jong, SIMPLS: An alternative approach to partial least squares regression, Chemom. Intell. Lab. Syst. 18 (1993) 251–263. doi:10.1016/0169-7439(93)85002-X.

[28] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta. 185 (1986) 1–17. doi:10.1016/0003-2670(86)80028-9.

Chapter IV

Variable selection for classification

Abstract

New methods for variable selection in classification using ordered predictors selection (OPS) are proposed in this work. The OPS method was initially designed to select variables in multivariate regression, and it has been successfully used for this purpose. However, a specific method that uses the OPS strategy to select variables in discriminant analysis has not yet been presented. OPS for feature selection in the discriminant analysis (OPSDA) selects the most predictive variables for each class using a cross-validation procedure and a window search based on partial least squares for discriminant analysis (PLS-DA). The classification model based on OPSDA variable selection can be performed in two different ways. The first strategy combines the set of selected variables for each class in a unique vector to build a single classification model. The second one uses the set of selected variables for each class individually to build classification models to predict each class at a time. In this work, PLS-DA and support vector machines for discriminant analysis (SVM-DA) models were built and compared. The new OPSDA methods were applied on three datasets, with different types of predictors, and the results were compared with the models using all variables. The OPSDA provided the best set of selected variables to build more predictive models using both classification methods, PLS-DA and SVM-DA. Therefore, OPSDA proved to be efficient to select variables in different types of datasets, and it could be applied to classification problems independent of the number of classes.

Keywords: Variable selection, classification, PLS-DA, SVM-DA.

Introduction

Pattern recognition is one of the major areas of chemometrics related to analytical chemistry, and it is widely used in chemical problem solving [1]. In this field, there are two approaches, depending on what is known a priori. When no information about samples grouping is required, the methods are called unsupervised pattern recognition, such as principal component analysis (PCA), hierarchical cluster analysis (HCA), and Kohonen networks. On the other hand, when the group assignments are known (classes), the methods are named supervised pattern recognition or classification methods. From this, there are a variety of methods available, such as soft independent modeling of class analogy (SIMCA), *k*-nearest neighbor (*k*-NN), decision trees, support vector machines for discriminant analysis (SVM-DA), linear discriminant analysis (LDA), and partial least squares for discriminant analysis (PLS-DA).

PLS-DA is a variant of partial least squares (PLS) regression and has often been applied to classification problems in many research fields [2–5]. A large number of variables in a dataset is a common issue in modern chemometrics and, sometimes, most of them are irrelevant for interpretation [6,7]. For this reason, there are several approaches to search for the most significant variables in supervised classification methods. Variable selection method using intra and inter-variance was applied in mid-infrared spectroscopy to build PLS-DA and SIMCA models [8]. Genetic algorithm (GA) [9] was used to select variables in PLS-DA for nuclear magnetic resonance spectroscopy [10,11] and terahertz spectroscopy [12]. Also, GA was applied combined with LDA [13] in infrared spectroscopy. Uninformative variable elimination (UVE) [14], variable importance in projection (VIP) [15], and significance multivariate correlation (SMC) [16] were used as feature selection with PLS-DA in chromatographic profiles [17]. SVM recursive feature elimination (SVM-RFE) [18] was used to select variables in mass spectrometry dataset [19]. Besides, VIP scores is a common method used to select variables in PLS-DA [19–25], LDA [19], and SIMCA [21].

Another variable selection method very used to built LDA models is the successive projections algorithm (SPA) [26–30]. Further, recursive weighted PLS (rPLS) [23,31], equidistant combination multiple linear regression (ECMLR) [32,33], interval PLS (*i*PLS) [11,12,34], and competitive adaptive reweighted sampling (CARS)[35] were used to select variables combined with PLS-DA.

The ordered predictors selection (OPS) [36,37] is a method to select variables in multivariate regression with the potential to be adapted for discriminant analysis. This method has the advantage to be simple, fast, and efficient for the selection of relevant predictors types [38–41]. A simple adaptation of the OPS algorithm for multivariate regression was applied to build classification models of two classes in some studies [38,42–44]. These results confirm the potential of the OPS algorithm to improve classification models. However, a specific algorithm, which uses the strategy used by OPS to select variables in the discriminant analysis, has not yet been presented.

Therefore, this study proposes the development of the ordered predictors selection for discriminant analysis (OPSDA), a method to select variables for classification methods. Then, three OPSDA approaches were proposed. The first one, automatic OPS for discriminant analysis (*autoOPSDA*), performs all calculations using either or both informative vectors and its combinations. The second one, called feedback OPS for discriminant analysis (*feedOPSDA*), wherein the pre-selected variables would go through a new selection run. The last one, interval OPS for discriminant analysis (*i*OPSDA), apply the OPSDA in subdivisions of the full array. Besides, the approaches were applied to different types of datasets with two, three, and four classes, with the proposal to demonstrate the application of OPSDA method in different data structures and multiclass classification problems.

Theory background

This section will describe the PLS-DA method, chosen as the classification method to be used in OPSDA, the parameters of model evaluation, and OPSDA strategies for variable selection.

Partial Least Squares for Discriminant Analysis (PLS-DA)

PLS-DA is a classification method based on the PLS approach [45]. This supervised pattern recognition method uses the **X** matrix of independent variables with a **y** vector of dependent dummy variables to build classification models. In the two-class case, usually, the values of the dependent variable are coded 1 for one class and 0 for the other class. In case of more than two classes, more than one column of **y** is obtained, resulting in a **Y** matrix, and a PLS2 algorithm is used. Figure IV.1 shows the **X** matrix divided by colors, each representing a class that is given a correspondent region in the Class vector.

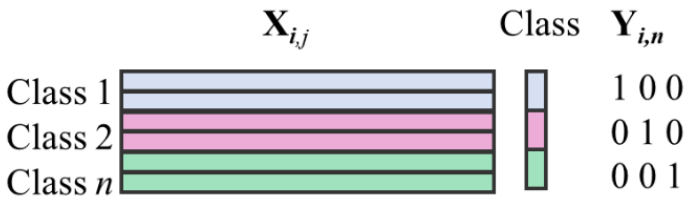

Figure IV.1. Scheme of **X** matrix (*i* samples and *j* variables), Class vector, and **Y** binary matrix (*i* samples and *n* classes) in PLS-DA.

This class vector can be rearranged in a binary matrix, where each column represents a class. In the first column of the **Y** matrix, ones are assigned in the rows corresponding to class 1 and zeros for the other classes' rows. In the second column, ones are assigned for the rows corresponding to class 2 and zeros for the other classes' rows, and so on.

In the PLS-DA model prediction, values close to zero and one are obtained, and this indicates that either the sample is or not in the modeled class, respectively. In practice, a threshold is determined by using a normal probability density function and Bayes decision [46], above which the sample belongs to the class and not belongs to the class if below.

Model Evaluation

Classification models were evaluated using the statistical parameters sensitivity and specificity. These parameters of decision making help in the reduction of the subjectivity in the performance evaluation. The sensitivity or true positive rate (equation (IV.1)) is the percentage of samples of each class accepted by the class model, and specificity or true negative rate (equation (IV.2)), is the percentage of samples of the other classes correctly rejected by the class model [45].

$$Sensitivity = \frac{TP}{(TP + FN)} \qquad \textbf{(IV.1)}$$

$$Specificity = \frac{TN}{(TN + FP)} \qquad \textbf{(IV.2)}$$

where *TP* is true positive (number of samples that belongs to the class classified as belonging to class), *FN* is false negative (number of samples that belongs to class not classified as belonging to class), *TN* is true negative (number of samples that does not belong to the class classified as not belonging to class), and *FP* is false positive (number of samples that does not belong to the class classified as belonging to class). Figure IV.2 illustrates these concepts.

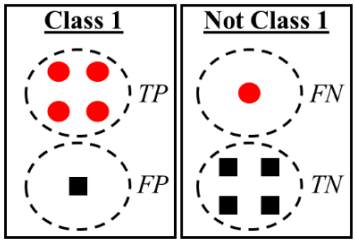

Figure IV.2. Parameters of decision making in classification models. *TP*: true positive; *FP*: false positive; *FN*: false negative; *TN*: true negative. (●) samples of class 1 and (■) samples of other class.

In addition, the classification error is also used to evaluate supervised pattern recognition models. Equation (IV.3) shows how to calculate this parameter.

$$Error = \frac{FP + FN}{(TP + TN + FP + FN)} \qquad \textbf{(IV.3)}$$

An ideal classification will have 1 for both sensitivity and specificity and error 0. The sensitivity and specificity decrease when false negatives or false positives occur, respectively. The increase in error depends on false negatives and positives. So, the desired error value is reached only when the maximum of both sensitivity and specificity are achieved.

Ordered Predictors Selection for Discriminant Analysis (OPSDA)

Briefly, this method is based on obtaining informative vectors containing information about the best response variables to increase the sensitivity and specificity. The models are obtained by sorting the data matrix in descending order of vector magnitude. Then, models are evaluated, and those with the best quality parameters of cross-validation are selected [36,37]. The model evaluation is performed using the classification error, where the variable set that provides the smallest classification error is chosen. OPSDA was designed to select variables based on cross-validation procedure building PLS-DA classification models. Even though the variables were selected using this approach, a PLS-DA model does not need to be built with these selected variables, another classification method, such as SVM-DA, *k*-NN, or LDA could be used to study these best variables.

In OPSDA method, it is possible to apply the variable selection using *auto*OPSDA, *feed*OPSDA, and *i*OPSDA. In *auto*OPSDA, all vector calculations are performed automatically using either or both informative vectors and its combinations, and the best one is chosen. Details about these vector calculations can be found elsewhere [36,37]. The best informative vector is

chosen based on the lowest classification error of cross-validation (ErrorCV), aiming to find the best variable subset to build predictive classification models.

The *feed*OPSDA apply feedback in OPSDA algorithm, wherein pre-selected variables can return to a new selection run until achieving the smallest ErrorCV. The ideal value of this parameter is zero, which indicate no misclassification. Convergence is achieved when in two consecutive selection runs, relative differences (*rd*) (equation (IV.4)) between error values are smaller than a value defined or when the error value increases instead of decrease. Thus, the previous loop is taken as the best selection.

$$rd = \frac{ErrorCV_n - ErrorCV_{n-1}}{ErrorCV_n} \qquad \textbf{(IV.4)}$$

where *n* is the number of selection runs. Besides that, the maximum number of selection runs can be defined, and so, the last loop is considered as the best selection. It is essential to highlight that in each selection run, the *auto*OPSDA is applied to provide the best informative vector.

The *i*OPSDA is applied in subdivisions of the full array. First, the number of variables in each interval is defined. The entire array is subdivided, and the variable selection is performed. In each interval, the *auto*OPSDA or the *feed*OPSDA is applied. This step is mainly used to reduce the number of variables in each interval. Then, the intervals comprising only the selected variables will be merged into a new matrix, and a new variable selection will take place considering the predetermined window and increment. The *auto*OPSDA (or *feed*OPSDA) is applied in the new array, and the best set of variables is reached.

The OPSDA perform the variable selection according to Figure IV.3. The feature selection is performed using the **X** matrix and each column of binary matrix **Y** at a time. Then, for each class is obtained a set of selected variables (varsel), which can be used in two different ways. The multiclass strategy (Figure IV.3A) combines the set of selected variables for each class (varsel 1 + varsel 2 + … + varsel n) in a unique vector to build a single classification model. If this combined vector has repeated variables, *i.e.*, the same variable was selected for more than one class, it would be counted just once (elimination of repeated variables), and a global set of selected variables is obtained. The one class-*vs*-all strategy (Figure IV.3B) uses each set of selected variables individually to build classification models to predict whether the given sample belongs or not to the given class. With *n* classes, in one class-*vs*-all, *n* classification models are built, except for two classes problem, where if the sample does not belong to one class, it belongs to another.

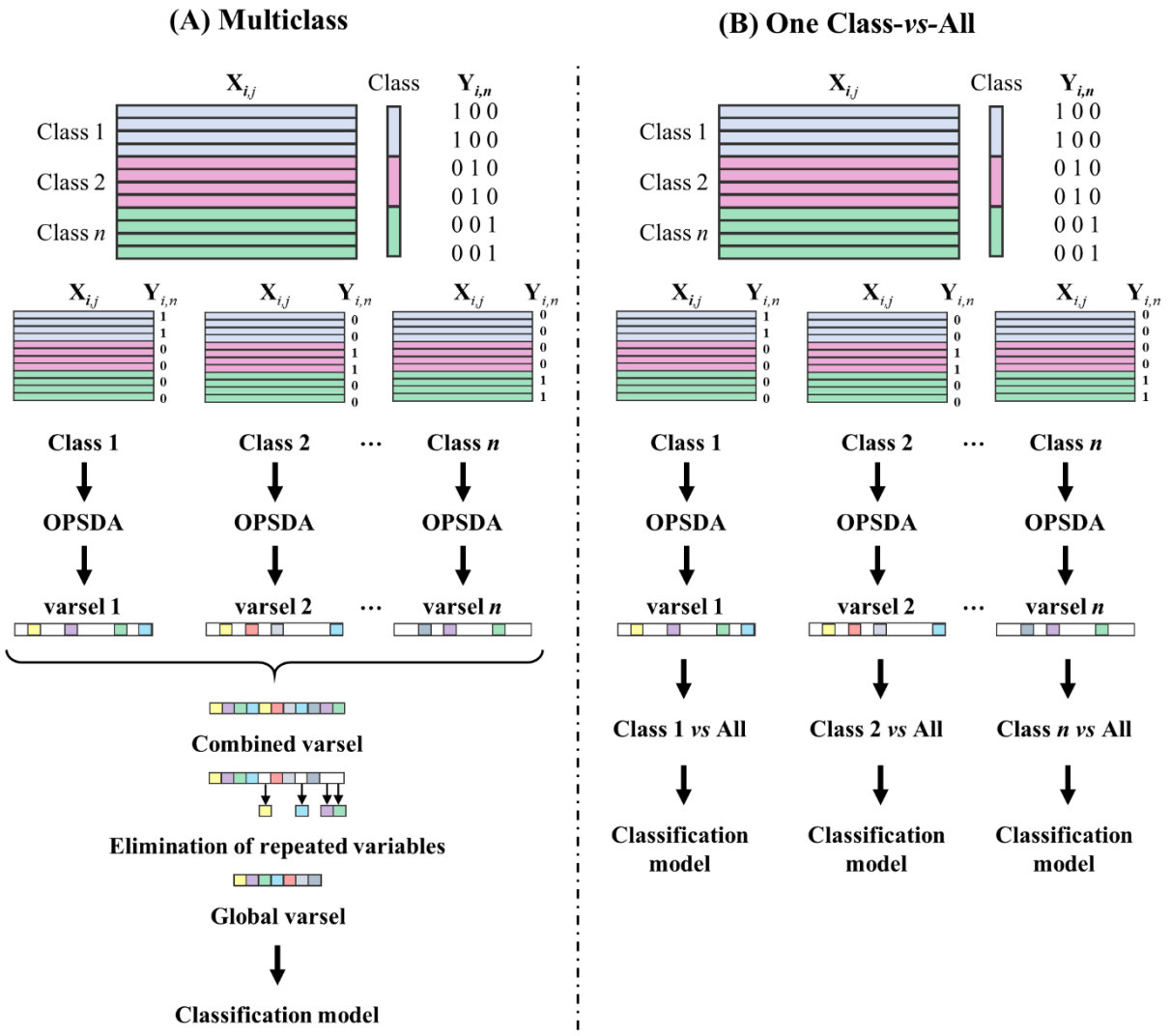

Figure IV.3. Scheme of (A) multiclass and (B) one-vs-all strategies to obtaining selected variables for classification models using OPSDA. **X** matrix (*i* samples and *j* variables), **Y** binary matrix (*i* samples and *n* classes), and varsel (the set of selected variables).

Experimental

Near-infrared spectroscopy, X-ray diffraction, and mass spectrometry datasets were used to build classification models with variables selected using OPSDA. The OPSDA were applied using 10 points window and increments of 5 variables, 100% of the variables were tested, random cross-validation was applied, with splits set to 10% of **X** matrix rows. In *feed*OPSDA, the convergence criteria were 2% as the minimum difference between two consecutive ErrorCVs and ten as the maximum number of loops. In *i*OPSDA, when the option to run the selection using *feed*OPSDA was used, the convergence criteria were the same as those used in *feed*OPSDA. The full matrix was divided into intervals of 10% of its size, limited to at least fifty variables. Additionally, *h*OPS, the number of latent variables used to calculate informative

vectors (more information can be found elsewhere [37]), was calculated for each interval in *i*OPSDA.

The full datasets were split into training and test sets using the Kennard-Stone [50] algorithm separately for each class to maintain classes proportion. The *auto*OPSDA, *feed*OPSDA, and *i*OPSDA methods using multiclass and one class-*vs*-all strategies were applied to each dataset testing all informative vectors. The OPSDA method that provided the best set of selected variables for each dataset was chosen to build classification models using the PLS-DA and SVM-DA.

The SVM-DA is primarily a classifier method that performs classification tasks by building hyperplanes in a multidimensional space that separates cases of different class labels [46,47]. An iterative training algorithm is used to minimize an error function to build an optimal hyperplane. In the Least Squares – Support Vector Machines (LS-SVM) [48] algorithm used in this work, a cost function is minimized, with the optimization of γ parameter. This is the regulation parameter of the machine learning algorithms that work to avoid overfitting. In nonlinear classification, kernel functions are used to perform a mapping of the data points (samples) in a new space of higher dimension, making them linearly separable. The radial basis function (RBF) is the most popular kernel function, and it depends on the σ^2 tuning. This parameter changes the degree of nonlinearity, which can be modeled. With an increase of σ^2, the kernel tends to find a linear solution [48,49]. Both parameters, γ and σ^2 were optimized for each classification model.

Four classification models were built for each dataset: PLS-DA and SVM-DA models using all variables (full models) and PLS-DA and SVM-DA models using selected variables by the best OPSDA. The algorithms for OPSDA and PLS-DA were written as *.m* function in MATLAB 2019a (Math Works, Natick, USA) and they are entirely authorial. The LS-SVM MATLAB/C Toolbox (available in https://www.esat.kuleuven.be/sista/lssvmlab/) were used to build SVM-DA models. All calculations were performed in MATLAB environment.

Near-Infrared Spectroscopy (NIR)

This dataset was composed by NIR spectra of sugarcane stalk samples supplied by the germplasm bank of the Sugarcane Genetic Breeding Program (PMGCA) from Universidade Federal de Viçosa, Viçosa, Minas Gerais (MG), Brazil. Reflectance NIR spectra were collected using a Fourier transform near-infrared (FT-NIR) spectrometer (Thermo Scientific Antaris II). Stalk scans were the average result of 32 scans measured with 4 cm^{-1} resolution over the

wavenumber range 10000 – 4000 cm^{-1}. These spectra were obtained in reflectance mode as log (1/R), where R is the collected reflectance.

Before building the classification models, the spectra were pre-treated using the second derivative, followed by mean centering. The training and test sets consisted of 139 and 20 samples, respectively. Stalk samples were classified according to a range of fiber content determined by a reference method. Samples were divided into two classes: the first one (class A) has low fiber content (7.000 to 13.400 %) and the second (class B) high fiber content (13.401 to 20.000 %).

X-Ray Diffraction (XRD)

This dataset consists of sugarcane stalk genotypes also supplied by PMGCA. Although this dataset and the previous one are both composed of stalks, they are from different samples. Pellets of sugarcane stalks were prepared in order to obtain XRD diffractograms. Briefly, stalk samples were grounded using a knife mill and 80-mesh sieve. The pellets were prepared using a press, transferring 0.40 g of powdered material to the sample set and applying 8.0 tons cm^{-1}. XRD measurements were performed in an X-ray diffractometer D8Discover (Bruker) equipped with Cu radiation source with wavelength 1.5418 Å (voltage of 40 kV and current of 40 mA). The θ-2θ scans were performed in the range of 10–40 degrees with steps of 0.1.

Second and first derivatives followed by autoscaling were performed in the diffractograms. Dataset was split into 21 and 9 samples for training and test sets, respectively. The classification model was built with three different classes of sugarcane: susceptible (class S), intermediate (class I) and resistant (class R) to sugarcane borer.

Mass Spectrometry (MS)

Pereira et al. [38] designed this dataset for the study of American lager beers from four different breweries. The mass spectra were obtained by a Thermo Fisher LCQ FLEET (San Jose, CA, USA) ion trap mass spectrometer in the range of 100 to 1800 m/z.

The beer mass spectra were transformed using the first derivative and pre-processed by mean centering. Although the first derivative is not indicated for mass spectra data, it expressively reduced the uninformative variables without shifting the relevant signal. Besides, the data presented an apparent continuous profile, which justifies the use of the first derivative. The samples were split into 93 and 48 for the training and test sets, respectively. The classes were assigned according to each brewery as follow: 72 beers of Anheuser-Busch InBev (class

A), 28 beers of Brazil Kirin (class B), 27 beers of Petropolis Group (class C), and 14 samples of Krill (class D). Details about this dataset can be found elsewhere [38].

Results and Discussion

Near-Infrared Spectroscopy (NIR)

In a two-class classification problem, the optimization of only one class is required. Then, the OPSDA methods were applied only once, *i.e.*, using one column of the **Y** variable. For NIR dataset, the *feedOPSDA* algorithm achieved the best result using REG as the informative vector in all selection runs. Initially, the data array was composed of 3113 wavenumbers. The number of selected informative wavenumbers decreases in each loop. Convergence was reached with seven loops (Figure IV.4), and a set of variables was selected by REG vector in loop six.

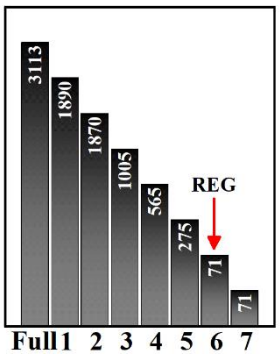

Figure IV.4. Number of selection runs in *feedOPSDA* in NIR dataset. The red arrow indicates the loop where the convergence was attained, and a set of selected variables was obtained.

Classification results using all variables (full) and selected variables using OPSDA-PLS-DA (OPSDA*pls*) and OPSDA-SVM-DA (OPSDA*svm*) are shown in Table IV.1. OPSDA*pls* and OPSDA*svm* models showed a significant improvement in comparison to full-PLS-DA and full-SVM-DA, respectively. Besides that, for both classes, these models proved excellent predictability (Specificity (Pred.) = 1).

Figure IV.5 shows the **Y** predicted values in the cross-validation of the full-PLS-DA model (Figure IV.5A) and OPSDA*pls* (Figure IV.5D). The cross-validation results are shown to demonstrate the improvement of the ErrorCV, which is the parameter used to select variables by OPSDA. Ideally, all red circle samples (class A) should be predicted above of dashed black line (threshold). In Figure IV.5A, full-PLS-DA model did not provide good separation in cross-validation procedure, while OPSDA*pls* model (Figure IV.5D) could predict the samples more correctly, with a significant decrease of ErrorCV. This model has only two false-positive

samples and one false negative for class A. Thus, the model that used selected variables was considered more efficient and robust.

Table IV.1. Statistical parameters of PLS-DA and SVM-DA models using all variables and selected variables by *feed*OPSDA for NIR dataset.

	PLS-DA		*SVM-DA*	
	Full	*OPS*	*Full*	*OPS*
			Class A	
nvars	3113	71	3113	71
nlvs / (γ,σ^2)	5	5	(29, 456085)	(5552, 3643)
Sensitivity (Training)	1.000	1.000	0.692	1.000
Sensitivity (Test)	0.875	1.000	0.375	1.000
Specificity (Training)	1.000	1.000	0.851	1.000
Specificity (Test)	0.917	1.000	0.917	1.000
Error (Training)	0.000	0.000	0.223	0.000
Error (Test)	0.100	0.000	0.300	0.000
			Class B	
nvars	3113	71	3113	71
nlvs / (γ,σ^2)	5	5	(29, 456085)	(5552, 3643)
Sensitivity (Training)	1.000	1.000	0.851	1.000
Sensitivity (Test)	0.917	1.000	0.917	1.000
Specificity (Training)	1.000	1.000	0.692	1.000
Specificity (Test)	0.875	1.000	0.375	1.000
Error (Training)	0.000	0.000	0.223	0.000

nvars: number of variables; *nlvs*: number of latent variables; γ and σ^2: optimization parameters of SVM-DA.

Figure IV.5B and IV.5C show the class prediction of the full-PLS-DA and full-SVM-DA models, respectively, and Figure IV.5E and IV.5F, the class prediction of the OPSDA*pls* and OPSDA*svm* models, respectively. Ideally, all red circle samples (class A) should be predicted as class A, *i.e.*, appearing in the bottom of the figure. In the opposite, all black squares (class B) should appear in the top of the figure, *i.e.*, predicted as class B. In full-PLS-DA model, the training set is correctly classified, but in the test set, there is misclassification in both classes. In full-SVM-DA, both training and test sets presented classification errors. In OPSDA*pls* and

OPSDA*svm* models, no misclassification is observed in both sets, training and test, showing maximum sensitivity and specificity (Table IV.1).

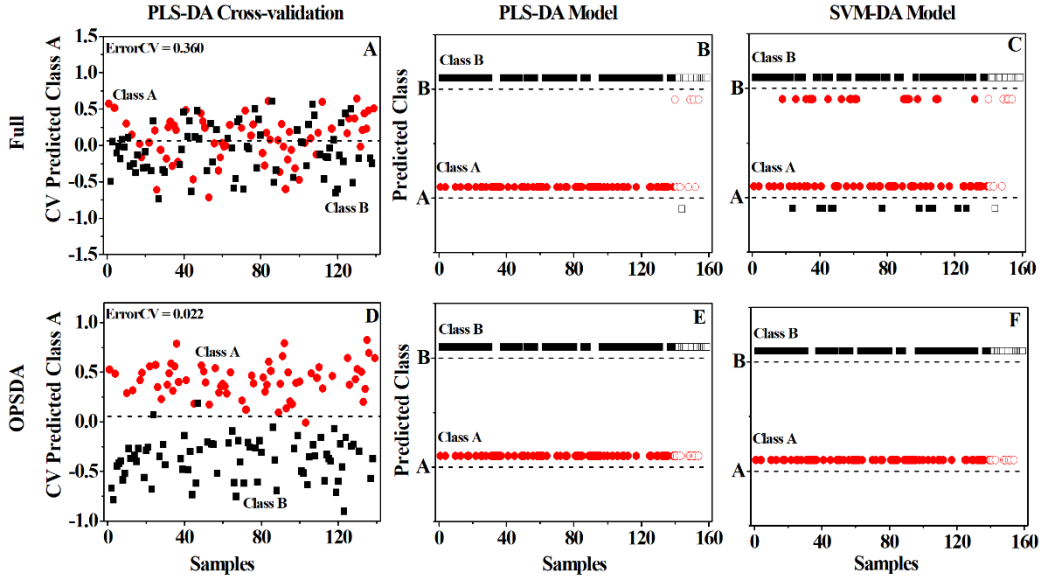

Figure IV.5. Cross-validation **Y** prediction values for NIR dataset in full-PLS-DA (A) and OPSDA*pls* (D). Classification by full-PLS-DA (B), OPSDA*pls* (C), full-SVM-DA (E), and OPSDA*svm* (F) models. Filled red circles (●) and empty red circles (○) are samples of class A. Filled black squares (■) and empty black squares (□) are samples of class B. Filled shapes refer to cross-validation samples, and the empty ones refer to test samples. In (A) and (D) the horizontal dashed black line indicates the threshold selected in PLS-DA method. In (B), (C), (E) and (F), indicates simply a class mark. ErrorCV: classification error in cross-validation.

In Figure IV.6A, it is presented the selected variables in NIR dataset by using the multiclass strategy. The variable selection occurred where well-defined regions are observed. Therefore, using the selected variables by OPSDA and NIR spectra was possible to classify correctly stalk samples of sugarcane genotypes in high (13.401 to 20.000 %) or low (7.000 to 13.400 %) fiber content.

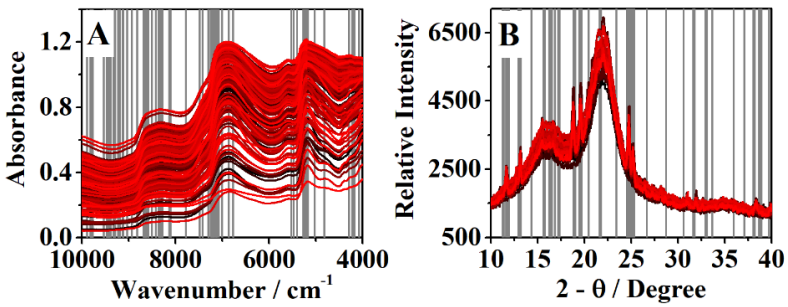

Figure IV.6. (A) NIR spectra and (B) XRD diffractograms with global selected variables by using the *feed*OPSDA with the multiclass strategy.

X-Ray Diffraction (XRD)

In Figure IV.7, the result of *feedOPSDA* algorithm is presented, which provided the best result using REG × SQR as the informative vector for all classes. The full matrix had 601 variables. Convergence was reached with two, three, and two loops for class S, I and R, respectively. So, a set of variables was selected for each class by REG ×SQR vector in loop one (class S), two (class I), and one (class R). For classes S and R, the best set of informative variables were obtained in the first selection run, *i.e.*, basically, the *autoOPSDA* algorithm was used to achieve these results. After that, the selected variables are combined in a multiclass strategy, and the global set of variables are obtained.

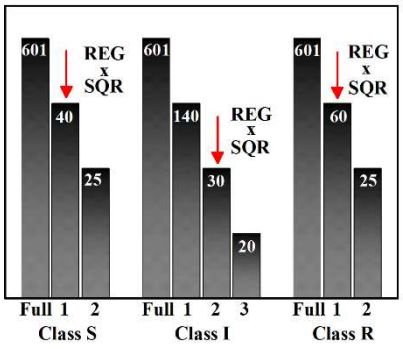

Figure IV.7. Number of selection runs in *feedOPSDA* for each class in XRD dataset. The red arrow indicates the loop where the convergence was attained, and the set of selected variables was obtained.

PLS-DA and SVM-DA results are shown in Table IV.2 and Figure IV.8. In each PLS-DA cross-validation plot, all red circle samples should be predicted above the dashed black line (threshold). For all classes, the ErrorCV significantly decreases with variable selection, reaching a higher sensitivity and specificity of cross-validation. In PLS-DA (Figure 8D and 8I) and SVM-DA (Figure IV.8E and IV.8J) model plots, ideally, all black diamond samples (class S) should be predicted as class S, *i.e.*, appearing in the bottom of the figure; all red diamond samples (class I) should be predicted as class I (the middle of the figure); and all blue diamond samples (class R) should appear in the top of the figure being classified as class R. In full models, the training set is correctly classified, but in the test set, there is misclassification in both models. In full-SVM-DA, all samples were predicted in the class R, showing zero sensitivity for classes S and I. In OPSDA*pls*, no misclassification is observed in both sets, training and test. However, in OPSDA*svm*, the test set still presented classification errors, with an increase of sensitivity for class I.

Table IV.2. Statistical parameters of PLS-DA and SVM-DA models using all variables and selected variables by *feed*OPSDA for XRD dataset.

	PLS-DA		*SVM-DA*	
	Full	*OPSDA*	*Full*	*OPSDA*
			Class Suscetible (S)	
nvars	3	3	(6.55, 1.23)	(5.60, 1.32)
nlvs / (γ,σ^2)	601	63	601	63
Sensitivity (Training)	1.000	1.000	1.000	1.000
Sensitivity (Test)	0.667	1.000	0.000	0.000
Specificity (Training)	1.000	1.000	1.000	1.000
Specificity (Test)	0.667	1.000	1.000	1.000
Error (Training)	0.000	0.000	0.000	0.000
Error (Test)	0.333	0.000	0.333	0.333
			Class Intermediate (I)	
nvars	3	3	(16, 2.48)	(161, 241)
nlvs / (γ,σ^2)	601	63	601	63
Sensitivity (Training)	1.000	1.000	1.000	1.000
Sensitivity (Test)	0.667	1.000	0.000	1.000
Specificity (Training)	1.000	1.000	1.000	1.000
Specificity (Test)	0.500	1.000	1.000	0.333
Error (Training)	0.000	0.000	0.000	0.000
			Class Resistant (R)	
nvars	3	3	(101, 729)	(4.07, 159)
nlvs / (γ,σ^2)	601	63	601	63
Sensitivity (Training)	1.000	1.000	1.000	1.000
Sensitivity (Test)	0.000	1.000	1.000	0.667
Specificity (Training)	1.000	1.000	1.000	1.000
Specificity (Test)	1.000	1.000	0.000	1.000
Error (Training)	0.000	0.000	0.000	0.000
nvars	0.333	0.000	0.667	0.111

nvars: number of variables; *nlvs*: number of latent variables; γ and σ^2: optimization parameters of SVM-DA.

In Figure IV.6B, it is displayed the global selected variables used to build classification models. The selected variables provide a better class separation of sugarcane genotypes stalk samples in susceptible, intermediate, or resistant to sugarcane borer using XRD and OPSDA regarding the full model.

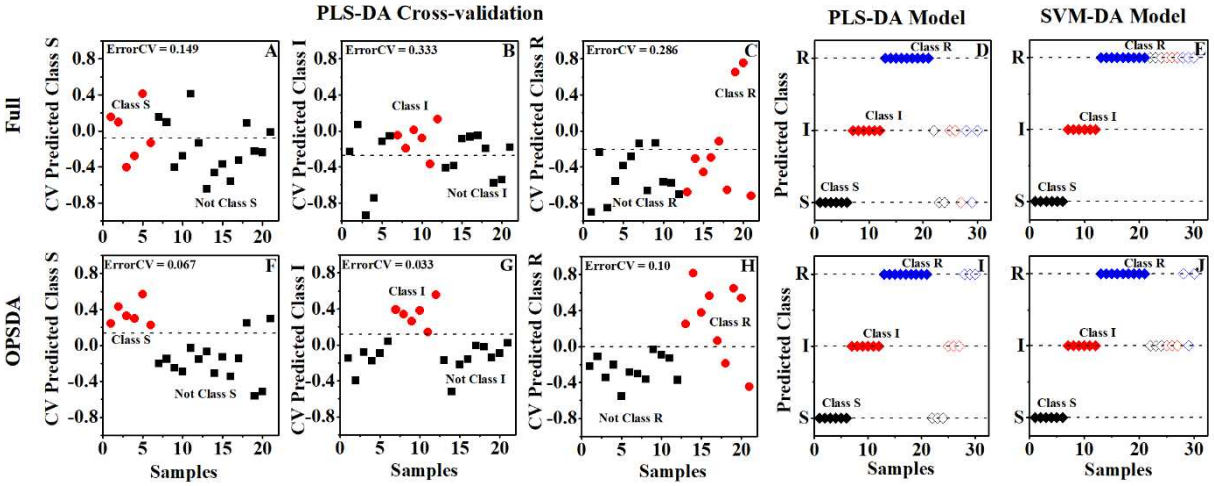

Figure IV.8. Cross-validation **Y** prediction values for XRD dataset in full-PLS-DA (A-C) and OPSDA*pls* (F-H). In each PLS-DA cross-validation plot, the red circles (●) are samples that refer to the class showed in the y-axis. Black squares (■) are samples that refer to other classes. In (A-C) and (F-H) the horizontal dashed black line indicates the threshold selected in PLS-DA method. Classification by full-PLS-DA (D), full-SVM-DA (E), OPSDA*pls* (I), and OPSDA*svm* (J) models. Black diamonds (♦ and ◊) represents class S, red diamonds (♦ and ◊) represents class I, and blue diamonds (♦ and ◊) represents class R. In (D-E) and (I-J), filled shapes refer to training samples and the empty ones refer to test samples; the horizontal dashed black line merely indicates a class mark.

Mass Spectrometry (MS)

In this dataset, the *feedOPSDA* algorithm also provided the best set of variables. Full data array was composed of 1697 variables or m/z values. The number of selected variables decreases in each loop until achieving a criterion. Convergence was reached with five, three, three, and five loops for class A, B, C, and D, respectively (Figure IV.9). Coincidentally, using the REG × SQR as the informative vector for all classes in all selection runs, a set of variables was selected for each class. For this dataset, four sets of variables were used to build individual models for each class using the one class-*vs*-all strategy. A classification model was built to predict whether a given sample belongs or not to class A. Another classification model was built to predict if a given sample belongs or not to class B, and so on for classes C and D.

PLS-DA and SVM-DA results for full and OPSDA models are shown in Table IV.3 and Figure IV.10. In each PLS-DA cross-validation plot, all red circle samples should be predicted

above the dashed black line (threshold). For class 2 and 4, good predictions were obtained with the full model (Figure IV.10B and IV.10D). OPSDA model provided the max of sensitivity and specificity in all classes, giving ErrorCV = 0 for all classes.

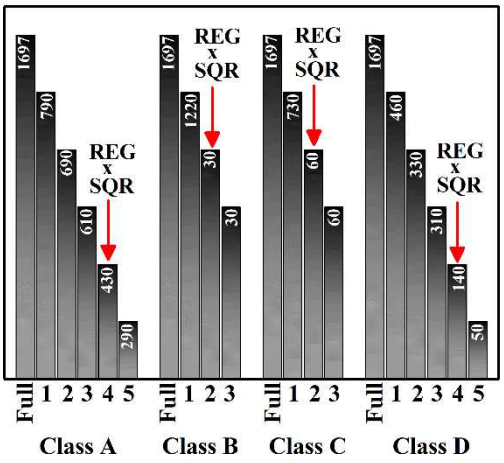

Figure IV.9. Number of selection runs in *feedOPSDA* for each class in XRD dataset. The red arrow indicates the loop where the convergence was attained, and the set of selected variables was obtained.

Figure IV.10I and IV.10K show the class prediction of the full-PLS-DA and full-SVM-DA models, respectively, and in Figure IV.10J and IV.10L, the class prediction of the OPSDA*pls* and OPSDA*svm* models, respectively. Ideally, all black diamond samples (class 1) should be predicted as class 1, *i.e.*, appearing in the bottom of the figure All red diamond samples (class 2) should be predicted as class 2 (the bottom middle of the figure) All blue diamond samples (class 3) should be predicted as class 3 (the top middle of the figure); All green diamond samples (class 4) should appear in the top of the figure being classified as class 4. From full models, all classes presented high sensitivity and specificity in training set in PLS-DA and SVM-DA models. In the test set, the full-SVM-DA also presented excellent predictability. On the other hand, for classes 1 and 3 in the test set, the sensitivity and specificity are lower than 1 in full-PLS-DA. From OPSDA models, for both PLS-DA and SVM-DA models, the sensitivity and specificity are ideal from training and test samples. In Figure IV.11, it is displayed the selected variables for each class provided by one class-*vs*-all strategy. The selected variables are different in each class, providing better class separation, and consequently, it was possible to classify beer samples correctly according to breweries using MS and OPSDA.

Table IV.3. Statistical parameters of PLS-DA and SVM-DA models using all variables and selected variables by *feed*OPSDA for MS dataset.

	PLS-DA		*SVM-DA*		*PLS-DA*		*SVM-DA*	
	Full	*OPSDA*	*Full*	*OPSDA*	*Full*	*OPSDA*	*Full*	*OPSDA*
			Class A				*Class B*	
nlvs / (γ, σ^2)	7	7	(422, 10581)	(323, 748)	7	7	(18, 10883)	(25, 2156)
nvars	1697	430	1697	430	1697	30	1697	30
Sensitivity (Training)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Sensitivity (Test)	0.917	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Specificity (Training)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Specificity (Test)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Error (Training)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Error (Test)	0.042	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			Class C				*Class D*	
nlvs / (γ, σ^2)	7	7	(131, 4372)	(270, 4769)	7	7	(1077, 4723)	(92, 539)
nvars	1697	60	1697	60	1697	140	1697	140
Sensitivity (Training)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Sensitivity (Test)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Specificity (Training)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Specificity (Test)	0.949	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Error (Training)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Error (Test)	0.042	0.000	0.000	0.000	0.000	0.000	0.000	0.000

nvars: number of variables; nlvs: number of latent variables; γ and σ^2: optimization parameters of SVM-DA.

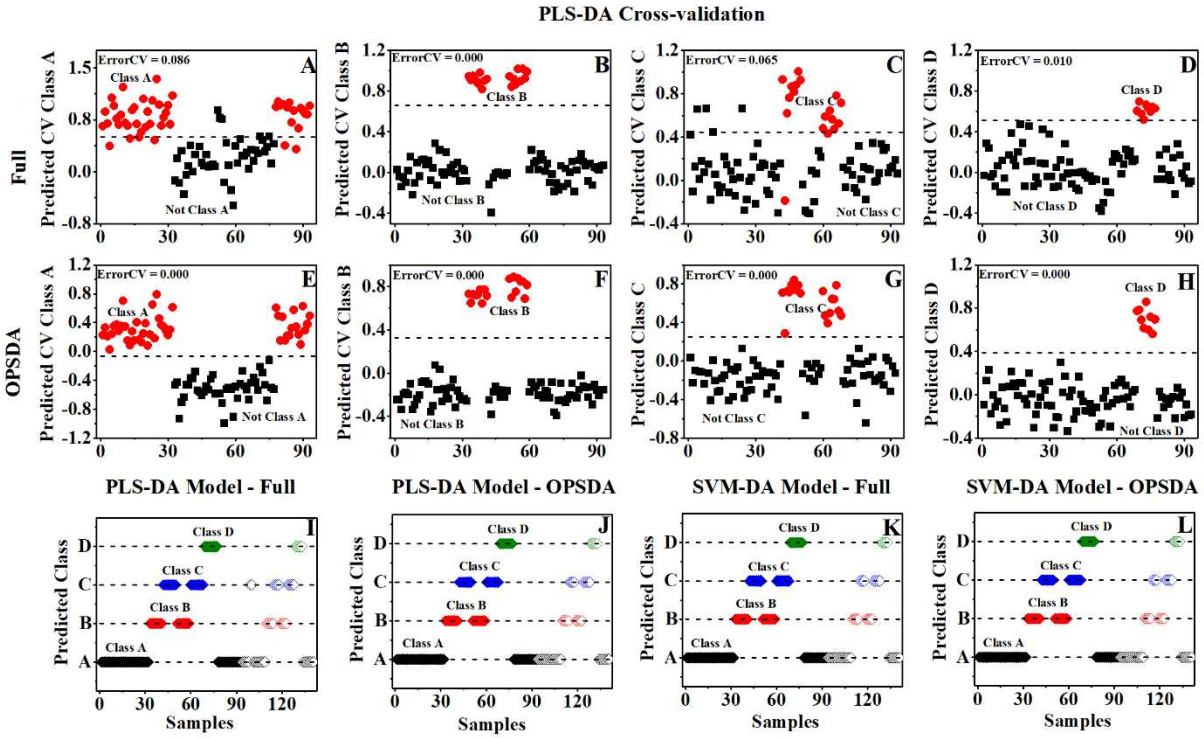

Figure IV.10. Cross-validation **Y** prediction values for MS dataset in full-PLS-DA (A-D) and OPSDA*pls* (E-H). In each PLS-DA cross-validation plot, the red circles (●) are samples that refer to the class showed in the y-axis. Black squares (■) are samples that refer to other classes. In (A-H) the horizontal dashed black line indicates the threshold selected in PLS-DA method. Classification by full-PLS-DA (I), full-SVM-DA (K), OPSDA*pls* (J), and OPSDA*svm* (L) models. Black diamonds (♦ and ◊) represents class A, red diamonds (♦ and ◊) represents class B, blue diamonds (♦ and ◊) represents class C, and green diamonds (♦ and ◊) represents class D. In (I-L), filled shapes refer to training samples and the empty ones refer to test samples; the horizontal dashed black line merely indicates a class mark.

In NIR dataset, the classification models were built to discriminate only two classes. In this case, OPSDA significantly enhances the predictability for both classes. When the number of classes increases, as in XRD and MS datasets, the model complexity also increases, and the gain of sensitivity can result in a loss of specificity, as happened in XRD. For this reason, in MS dataset with four classes, the one class-*vs*-all strategy was used to achieve the best classification results. The global results for MS dataset (results not shown) presented the gain of sensitivity for one class with the loss of specificity for another. Besides that, classification models built with selected variables for each class resulted in models more simple for interpretation. Therefore, in general, OPSDA algorithms were able to improve class predictions using either PLS-DA or SVM-DA methods.

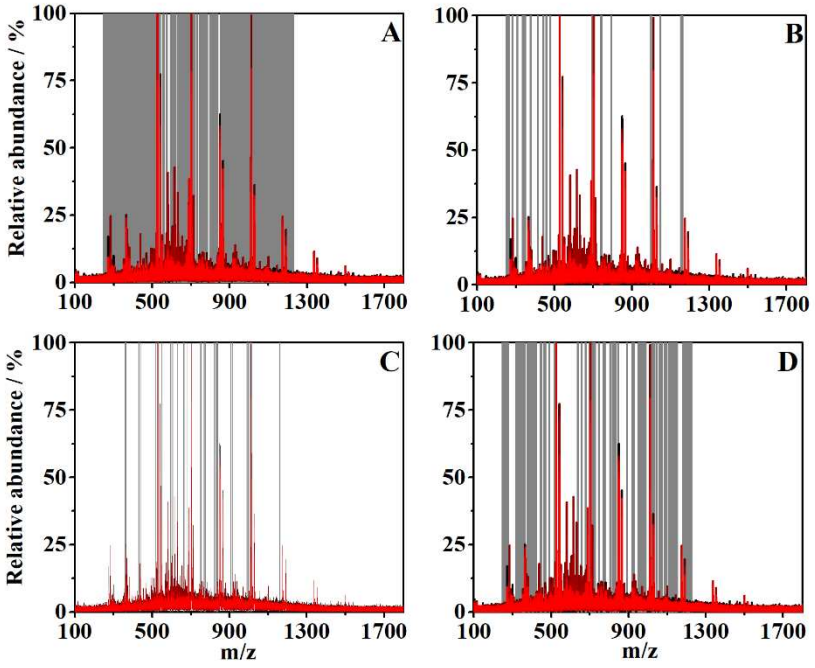

Figure IV.11. MS spectra with selected variables by OPSDA for class A (A), class B (B), class C (C), and class D (D).

Conclusion

The new OPSDA algorithms, *auto*OPSDA, *feed*OPSDA, *and i*OPSDA were applied to build classification models using PLS-DA and SVM-DA methods in NIR, XRD, and MS datasets. Overall, these new algorithms were able to enhance the model prediction power and reduce the number of variables significantly in all datasets. The OPSDA can provide a set of selected variables for each class separately, and this could make models simple for interpretations. Besides, the expansion of the OPS method to classification problems with two or more classes was successfully developed. Therefore, OPSDA could be applied to select variables to build classification models independent of the number of classes. In addition, selected variables by OPSDA can be used in discriminant analysis using different classification methods.

References

[1] P.K. Hopke, The evolution of chemometrics, Anal. Chim. Acta. 500 (2003) 365–377. doi:10.1016/S0003-2670(03)00944-9.

[2] L. Xie, Y. Ying, T. Ying, H. Yu, X. Fu, Discrimination of transgenic tomatoes based on visible/near-infrared spectra, Anal. Chim. Acta. 584 (2007) 379–384. doi:10.1016/j.aca.2006.11.071.

[3] J. Xing, W. Saeys, J. De Baerdemaeker, Combination of chemometric tools and image processing for bruise detection on apples, Comput. Electron. Agric. 56 (2007) 1–13.

doi:10.1016/j.compag.2006.12.002.
[4] P. Menesatti, A. Zanella, S. D'Andrea, C. Costa, G. Paglia, F. Pallottino, Supervised multivariate analysis of hyper-spectral NIR images to evaluate the starch index of apples, Food Bioprocess Technol. 2 (2009) 308–314. doi:10.1007/s11947-008-0120-8.
[5] R. Rodriguez-Bermudez, M. Lopez-Alonso, M. Miranda, R. Fouz, I. Orjales, C. Herrero-Latorre, Chemometric authentication of the organic status of milk on the basis of trace element content, Food Chem. 240 (2018) 686–693. doi:10.1016/j.foodchem.2017.08.011.
[6] T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K.-M. Myhr, O.M. Kvalheim, Discriminating Variable Test and Selectivity Ratio Plot: Quantitative Tools for Interpretation and Variable (Biomarker) Selection in Complex Spectral or Chromatographic Profiles, Anal. Chem. 81 (2009) 2581–2590. doi:10.1021/ac802514y.
[7] K. Wongravee, N. Heinrich, M. Holmboe, M.L. Schaefer, R.R. Reed, J. Trevejo, R.G. Brereton, Variable Selection Using Iterative Reformulation of Training Set Models for Discrimination of Samples: Application to Gas Chromatography/Mass Spectrometry of Mouse Urinary Metabolites, Anal. Chem. 81 (2009) 5204–5217. doi:10.1021/ac900251c.
[8] O. Galtier, O. Abbas, Y. Le Dreau, C. Rebufa, J. Kister, J. Artaud, N. Dupuy, Comparison of PLS1-DA, PLS2-DA and SIMCA for classification by origin of crude petroleum oils by MIR and virgin olive oils by NIR for different spectral regions, Vib. Spectrosc. 55 (2011) 132–140. doi:10.1016/j.vibspec.2010.09.012.
[9] R. Leardi, A. Lupiáñez González, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemom. Intell. Lab. Syst. 41 (1998) 195–207. doi:10.1016/S0169-7439(98)00051-3.
[10] Z. Ramadan, D. Jacobs, M. Grigorov, S. Kochhar, Metabolic profiling using principal component analysis, discriminant partial least squares, and genetic algorithms, Talanta. 68 (2006) 1683–1691. doi:https://doi.org/10.1016/j.talanta.2005.08.042.
[11] C.V. Di Anibal, M.P. Callao, I. Ruisánchez, 1H NMR variable selection approaches for classification. A case study: The determination of adulterated foodstuffs, Talanta. 86 (2011) 316–323. doi:https://doi.org/10.1016/j.talanta.2011.09.019.
[12] M. Yin, S. Tang, M. Tong, Identification of edible oils using terahertz spectroscopy combined with genetic algorithm and partial least squares discriminant analysis, Anal. Methods. 8 (2016) 2794–2798. doi:10.1039/c6ay00259e.
[13] A.C. Silva, L.F.B.L. Pontes, M.F. Pimentel, M.J.C. Pontes, Detection of adulteration in hydrated ethyl alcohol fuel using infrared spectroscopy and supervised pattern recognition methods, Talanta. 93 (2012) 129–134. doi:https://doi.org/10.1016/j.talanta.2012.01.060.
[14] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Elimination of uninformative variables for multivariate calibration., Anal. Chem. 68 (1996) 3851–3858. doi:10.1021/ac960321m.
[15] O.M. Kvalheim, R. Arneberg, O. Bleie, T. Rajalahti, A.K. Smilde, J.A. Westerhuis, Variable importance in latent variable regression models, J. Chemom. 28 (2014) 615–622. doi:10.1002/cem.2626.
[16] T.N. Tran, N.L. Afanador, L.M.C. Buydens, L. Blanchet, Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC), Chemom. Intell. Lab. Syst. 138 (2014) 153–160. doi:https://doi.org/10.1016/j.chemolab.2014.08.005.
[17] B. Krakowska, D. Custers, E. Deconinck, M. Daszykowski, The Monte Carlo validation framework for the discriminant partial least squares model extended with variable

selection methods applied to authenticity studies of Viagra® based on chromatographic impurity profiles, Analyst. 141 (2016) 1060–1070. doi:10.1039/C5AN01656H.

[18] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene Selection for Cancer Classification using Support Vector Machines, Mach. Learn. 46 (2002) 389–422. doi:10.1023/A:1012487302797.

[19] P.S. Gromski, Y. Xu, E. Correa, D.I. Ellis, M.L. Turner, R. Goodacre, A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data, Anal. Chim. Acta. 829 (2014) 1–8. doi:https://doi.org/10.1016/j.aca.2014.03.039.

[20] F.C. De Lucia, J.L. Gottfried, Influence of variable selection on partial least squares discriminant analysis models for explosive residue classification, Spectrochim. Acta Part B At. Spectrosc. 66 (2011) 122–128. doi:https://doi.org/10.1016/j.sab.2010.12.007.

[21] J.A.R. Teodoro, H. V Pereira, D.N. Correia, M.M. Sena, E. Piccin, R. Augusti, Forensic discrimination between authentic and counterfeit perfumes using paper spray mass spectrometry and multivariate supervised classification, Anal. Methods. 9 (2017) 4979–4987. doi:10.1039/C7AY01295K.

[22] M.A.B. Hedegaard, K.L. Cloyd, C.-M. Horejs, M.M. Stevens, Model based variable selection as a tool to highlight biological differences in Raman spectra of cells, Analyst. 139 (2014) 4629–4633. doi:10.1039/c4an00731j.

[23] G. Aliakbarzadeh, H. Parastar, H. Sereshti, Classification of gas chromatographic fingerprints of saffron using partial least squares discriminant analysis together with different variable selection methods, Chemom. Intell. Lab. Syst. 158 (2016) 165–173. doi:10.1016/j.chemolab.2016.09.002.

[24] G.R. Lloyd, N. Stone, Method for Identification of Spectral Targets in Discrete Frequency Infrared Spectroscopy for Clinical Diagnostics, Appl. Spectrosc. 69 (2015) 1066–1073. doi:10.1366/14-07677.

[25] H.E. Park, S.-Y. Lee, S.-H. Hyun, D.Y. Kim, P.J. Marriott, H.-K. Choi, Gas chromatography/mass spectrometry-based metabolic profiling and differentiation of ginseng roots according to cultivation age using variable selection., J. AOAC Int. 96 (2013) 1266–1272.

[26] S.F.C. Soares, A.A. Gomes, M.C.U. Araujo, A.R.G. Filho, R.K.H. Galvão, The successive projections algorithm, TrAC Trends Anal. Chem. 42 (2013) 84–98. doi:https://doi.org/10.1016/j.trac.2012.09.006.

[27] P.H.G.D. Diniz, A.A. Gomes, M.F. Pistonesi, B.S.F. Band, M.C.U. de Araújo, Simultaneous Classification of Teas According to Their Varieties and Geographical Origins by Using NIR Spectroscopy and SPA-LDA, Food Anal. Methods. 7 (2014). doi:10.1007/s12161-014-9809-7.

[28] V.E. de Almeida, G.B. da Costa, D.D.S. Fernandes, P.H.G.D. Diniz, D. Brandao, A.C. de Medeiros, G. Veras, Using color histograms and SPA-LDA to classify bacteria, Anal. Bioanal. Chem. 406 (2014) 5989–5995. doi:10.1007/s00216-014-8015-1.

[29] P.H.G.D. Diniz, M.F. Pistonesi, M.B. Alvarez, B.S. Fernandez Band, M.C. de Araujo, Simplified tea classification based on a reduced chemical composition profile via successive projections algorithm linear discriminant analysis (SPA-LDA), J. FOOD Compos. Anal. 39 (2015) 103–110. doi:10.1016/j.jfca.2014.11.012.

[30] U.T.C.P. Souto, M.F. Barbosa, H.V. Dantas, A.S. de Pontes, W. S. Lyra, P.H.G.D. Dias Diniz, M.C. de Araujo, E.C. da Silva, Screening for Coffee Adulteration Using Digital Images and SPA-LDA, Food Anal. Methods. 8 (2015) 1515–1521. doi:10.1007/s12161-014-0020-7.

[31] Å. Rinnan, M. Andersson, C. Ridder, S.B. Engelsen, Recursive weighted partial least

squares (rPLS): an efficient variable selection method using PLS, J. Chemom. 28 (2014) 439–447. doi:10.1002/cem.2582.

[32] T. Pan, M. Li, J. Chen, Selection method of quasi-continuous wavelength combination with applications to the near-infrared spectroscopic analysis of soil organic matter., Appl. Spectrosc. 68 (2014) 263–271. doi:10.1366/13-07088.

[33] L. Zhang, M. Sun, Z. Wang, H. Li, Y. Li, Z. Fu, Y. Guan, G. Li, L. Lin, Optimal wavelength selection for visible diffuse reflectance spectroscopy discriminating human and nonhuman blood species, Anal. Methods. 8 (2016) 381–385. doi:10.1039/C5AY02865E.

[34] R. Leardi, L. Nørgaard, Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, J. Chemom. 18 (2004) 486–497. doi:10.1002/cem.893.

[35] W. Fan, H. Li, Y. Shan, H. Lv, H. Zhang, Y. Liang, Classification of vinegar samples based on near infrared spectroscopy combined with wavelength selection, Anal. Methods. 3 (2011) 1872–1876. doi:10.1039/c1ay05101f.

[36] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression, J. Chemom. 23 (2009) 32–48. doi:10.1002/cem.1192.

[37] J. V. Roque, W. Cardoso, L.A. Peternelli, R.F. Teofilo, R.F. Teófilo, Comprehensive new approaches for variable selection using ordered predictors selection, Anal. Chim. Acta. (2019). doi:10.1016/j.aca.2019.05.039.

[38] H.V. Pereira, V.S. Amador, M.M. Sena, R. Augusti, E. Piccin, Paper spray mass spectrometry and PLS-DA improved by variable selection for the forensic discrimination of beers, Anal. Chim. Acta. 940 (2016) 104–112. doi:https://doi.org/10.1016/j.aca.2016.08.002.

[39] L.J. de Campos, E.B. de Melo, Modeling structure-activity relationships of prodiginines with antimalarial activity using GA/MLR and OPS/PLS., J. Mol. Graph. Model. 54 (2014) 19–31. doi:10.1016/j.jmgm.2014.08.004.

[40] G.A. Silva, D.A. Maretto, H.M. a Bolini, R.F. Teófilo, F. Augusto, R.J. Poppi, Correlation of quantitative sensorial descriptors and chromatographic signals of beer using multivariate calibration strategies, Food Chem. 134 (2012) 1673–1681. doi:10.1016/j.foodchem.2012.03.080.

[41] J. V Roque, L.A.S. Dias, R.F. Teófilo, Multivariate Calibration to Determine Phorbol Esters in Seeds of Jatropha curcas L. Using Near Infrared and Ultraviolet Spectroscopies, J. Braz. Chem. Soc. 28 (2017) 1506–1516. doi:10.21577/0103-5053.20160332.

[42] J.S. Ribeiro, F. Augusto, T.J.G. Salva, R.A. Thomaziello, M.M.C. Ferreira, Prediction of sensory properties of Brazilian Arabica roasted coffees by headspace solid phase microextraction-gas chromatography and partial least squares, Anal. Chim. Acta. 634 (2009) 172–179. doi:10.1016/j.aca.2008.12.028.

[43] K.L. Lang, I.T. Silva, V.R. Machado, L.A. Zimmermann, M.S.B. Caro, C.M.O. Simoes, E.P. Schenkel, F.J. Duran, L.S.C. Bernardes, E.B. de Melo, Multivariate SAR and QSAR of cucurbitacin derivatives as cytotoxic compounds in a human lung adenocarcinoma cell line, J. Mol. Graph. Model. 48 (2014) 70–79. doi:10.1016/j.jmgm.2013.12.004.

[44] L.A. Peternelli, M.H.P. Barbosa, J.V. Roque, R.F. Teofilo, Phenotypic classification of sugarcane from near infrared spectra obtained directly from stalk using ordered predictors selection and partial least squares-discriminant analysis, in: Proc. 18th Int. Conf. Near Infrared Spectrosc., IM Publications Open LLP, 2019: pp. 157–161. doi:10.1255/nir2017.157.

[45] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemom. 17 (2003)

166–173. doi:10.1002/cem.785.
[46] N.F. Pérez, J. Ferré, R. Boqué, Calculation of the reliability of classification in discriminant partial least-squares binary classification, Chemom. Intell. Lab. Syst. 95 (2009) 122–128. doi:10.1016/j.chemolab.2008.09.005.
[47] B. Üstün, W.J. Melssen, L.M.C. Buydens, Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel, Chemom. Intell. Lab. Syst. 81 (2006) 29–40. doi:10.1016/j.chemolab.2005.09.003.
[48] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, Berlin, Heidelberg, 1995.
[49] J.A.K. Suykens, J. Vandewalle, Least Squares Support Vector Machine Classifiers, Neural Process. Lett. 9 (1999) 293–300. doi:10.1023/A:1018628609742.
[50] R.W. Kennard, L.A. Stone, Computer Aided Design of Experiments, Technometrics. 11 (1969) 137–148. doi:10.1080/00401706.1969.10490666.

CONCLUDING REMARKS

New comprehensive and versatile OPS methods for regression and supervised pattern recognition were developed. The prediction performance of the new OPS strategies for regression outperformed the OPS*v*1. These new methods were applied to several types of datasets, and their external prediction was better in comparison to other variable selection methods. The choice of the variable selection method is a crucial step to be considered due to the significant differences of the algorithms running time. The dataset matrix dimensions have a more pronounced effect in running time, while the number of latent variables has a minor effect. Although the OPS methods were not the fastest algorithms, they provided the most predictive and interpretative models in comparison to other variable selection methods. Regarding the OPS methods for classification, the OPSDA algorithms select variables for each class separately, independent of the number of classes, which has the potential to make models simple for interpretation. Also, the variables selected by the OPSDA methods could be used for discriminant analysis using different supervised pattern recognition methods. Overall, the new OPS methods proved to be a universal and powerful method that significantly improved the prediction ability for regression and classification purposes.