# Variable Selection as a Non-Completely Decomposable Problem: A Case Study in Multivariate Calibration

Lauro C. M. de Paula
Federal University of Goiás
Goiânia, Goiás, Brazil
laurocassio@inf.ufg.br

Anderson S. Soares
Federal University of Goiás
Goiânia, Goiás, Brazil
anderson@inf.ufg.br

Telma W. Soares
Federal University of Goiás
Goiânia, Goiás, Brazil
telma@inf.ufg.br

Clarimar J. Coelho
Pontifical Catholic University of Goiás
Goiânia, Goiás, Brazil
clarimarc@gmail.com

## ABSTRACT

Variable selection is a procedure used to choose a subset of features in order to extract information from them. It has been widely used in multivariate calibration together with statistical techniques to build a model from which it is possible to be interpreted by users. Genetic algorithms (GAs) have been successfully utilized as a variable selection method in multivariate calibration models. However, GAs solve a problem by trying different decompositions, and the variable selection problem usually can not be properly decomposed when there is considerable correlation among variables. Consequently, GAs tend to lead to a poor variable selection performance if the variables interdepence is strong. This work comes from a doctoral thesis, which is still in development and aims to (not only) demonstrate that selecting variables in multivariate calibration is a non-completely decomposable problem. Based on the preliminary results, we are able to claim the viability of our initial hypothesis.

## KEYWORDS

Decomposability; Multivariate Calibration; Variable Selection; Genetic Algorithms.

**ACM Reference format:**
Lauro C. M. de Paula, Anderson S. Soares, Telma W. Soares, and Clarimar J. Coelho. 2017. Variable Selection as a Non-Completely Decomposable Problem: A Case Study in Multivariate Calibration. In *Proceedings of GECCO '17 Companion, Berlin, Germany, July 15-19, 2017,* 4 pages.
DOI: http://dx.doi.org/10.1145/3067695.3082494

## 1 INTRODUCTION

Multivariate calibration is a sub-area of study from chemometrics related to analytical chemistry. It determines a mathematical model which relates the data to a given property of interest (*e.g.*, protein, pharmaceutical ingredient, qualitative parameters) from known samples (*e.g.*, wheat, medicine, vegetable oils) in order to predict this property by selecting informative variables [5].

Variable selection is the procedure used to choose a subset of suitable features contained in a given data set. Selecting variables becomes important when the data set contains many redundant and irrelevant features. Such features usually do not provide distinguished knowledge and should be removed without incurring loss of information [3].

Given a set of explanatory variables in a matrix $\mathbf{X}$ and a set of response variables in a vector $\mathbf{y}$, the learning task can be synthesized to find $\mathbf{y} = f(\mathbf{X})$. Among the main statistical techniques used for performing the calibration process and obtaining mathematical models is the multiple linear regression (MLR) [5]. MLR is a statistical technique used to build models which describe the relationships among several informative variables [2]. Multivariate calibration utilizes MLR, which is a common tool between chemometrics and machine learning to construct the models.

In order to deal with larger and more complex datasets, the development of efficient variable selection methods becomes an increasingly important asset. Thus, several studies have proposed evolutionary algorithms (EAs) for the variable selection procedure in multivariate calibration. For instance, Xu *et al.* [13] presented a genetic algorithm (GA) for variable selection in visible and near-infrared spectra. Authors showed that the proposed GA can be used for industrial applications. Niazi and Leardi [6] published a review which covers the application of GAs in chemometrics. The goal was to show the main research fields of GAs applications together with providing a list of references on the subject. On the other hand, Paula *et al.* [8] demonstrated that one-point crossover used by standard GAs tends to cause the building blocks disruption, which usually leads to undesirable performance.

It is known that decomposable problems can be created by concatenating basis functions of a certain order [1]. For a problem to be decomposable, there must be no interaction between any two variables and each variable should be separately treated [11]. However, often in multivariate calibration there are considerable linear dependencies among decision variables from spectral data [2, 5]. According to Watson [12], a set of correlated variables is not decomposable.

This paper is based on a doctoral thesis being developed. Our hypothesis claims that spectral data in multivariate calibration may be considered as a non-completely decomposable problem due to the


*GECCO '17 Companion, Berlin, Germany*

DOI: http://dx.doi.org/10.1145/3067695.3082494

constant presence of data correlation. As a consequence, GAs tend to lead to an unsuitable variable selection performance since such problems may not be properly decomposed [8, 11]. Our examples and preliminary results point out the hypothesis feasibility.

Section 2 discusses about the main concepts regarding variable selection in multivariate calibration. Section 3 presents our proposed hypothesis. The materials and methods used to obtain the outcomes are described in Section 4. Results are discussed in Section 5. Finally, Section 6 provides the final conclusions and the next steps in the work.

## 2 BACKGROUND

A mathematical model can be obtained to measure the concentration level of a certain property of interest from the sample. Usually, reference values (previously yielded in laboratory) can be used to assess the model predictive ability. Such a mathematical model establishes the relationship between the properties measured by the spectrophotometer and the concentration of an analysed sample [5]. It can be used to provide the value of a quantity $\mathbf{y}$ based on values measured from a set of explanatory variables $\mathbf{X}_{cal} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$, and can be defined by Equation (1):

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + ... + \beta_k \mathbf{x}_k + \varepsilon, \tag{1}$$

where $\beta_0, \beta_1, ..., \beta_k$ are the regression coefficients to be determined, and $\varepsilon$ is a measure of random error.

In order to obtain the coefficients in Equation (1), one may use the multiple linear regression (MLR). MLR is a statistical technique used to build models that describe the relationships among several explanatory variables [2]. Equation (2) shows how those regression coefficients can be calculated:

$$\beta = (\mathbf{X}_{cal}^T \mathbf{X}_{cal})^{-1} \mathbf{X}_{cal}^T \mathbf{y}, \tag{2}$$

where $\mathbf{X}_{cal}$ is the $n \times k$ matrix of variables and observations from the calibration set, $\mathbf{y}$ is the vector of reference variables, and $\beta$ is the vector of regression coefficients.

Soon after calculating those coefficients, it becomes necessary to determine the predictive ability of MLR models. This may be achieved by comparing predictions with reference values for a test set and using statistical measures. The root mean square error of prediction (RMSEP) is a measure of the differences between values predicted by a model and the values actually observed. In the context of multivariate calibration, it is depicted as shown in Equation (3):

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{n}}, \tag{3}$$

where $\mathbf{y}$ is the reference values of the property of interest (which is attempted to be determined in the analysed sample), $n$ is the number of observations (number of rows of matrix $\mathbf{X}_{cal}$), and $\hat{\mathbf{y}}$ is the estimated value.

Based on the RMSEP value, it becomes possible to determine if the model has an adequate predictive ability or not. In general, the goal consists of obtaining a model with a considerably reduced error value. However, in order to achieve this goal, it often becomes necessary to deal with multicollinearity problem.

One of the main issues related to the calibration process is the recurrent presence of linear correlations among variables. The existence of linear correlation between two or more variables is a mathematical problem defined as multicollinearity [2]. Multicollinearity can be caused by the relationship among explanatory variables, and it is an undersirable attribute of the particular calibration set that has been collected. It can reduce the reliability of coefficients from estimated models [2]. In literature, it is possible to find many techniques to deal with multicollinearity [2]. Reducing the number of variables is a significant procedure which has been widely used.

Selecting a reduced set of informative (explanatory) variables becomes important to improve the efficiency of techniques used to construct MLR models. Hence, smaller RMSEP values can be achieved. In this sense, the use of variable selection methods has become important an approach to deal with multicollinearity [8].

## 3 PROPOSED HYPOTHESIS

Additively decomposable functions are one of the representations of a decomposable problem. Franz [11] states that decomposability is one of the reasons for the advantage of GAs performing over a random search. Assuming the variable selection procedure in multivariate calibration as a decomposable problem, an additively decomposable function can be defined as

$$f(\mathbf{X}_{n\times k}) = \sum_{i=1}^{N} f_{\mathbf{v}^i_{1\times m}}(\mathbf{S}^i_{n\times m}), \forall\ \mathbf{v}^i_{1\times m} \in \mathbf{V}_{N\times m}, \tag{4}$$

where $\mathbf{X}_{n\times k}$ is the matrix from Equation (2), $\mathbf{S}^i_{n\times m}$, $1 \leq m \leq k$, are $N$ different subsets of variables (columns) from $\mathbf{X}_{n\times k}$, $\mathbf{v}^i_{1\times m}$ is an individual from the population $\mathbf{V}_{N\times m}$ of a GA, and $N$ is the number of individuals in $\mathbf{V}_{N\times m}$. More specifically, every $\mathbf{v}^i_{1\times m}$ is an $1 \times m$ binary vector in which each element equals to 1 means the respective variable is to be selected. Otherwise, an element equals to 0 does not select any variable.

In this sense, the variable subsets could be divided into several smaller subsets and recombined by genetic operators to form new better individuals. However, a subset of informative variables may provide better outcomes than these variables divided into several subsets. Thus, such subset should not be split into different parts.

When it is not possible to break a problem into smaller subproblems or its pieces affect one another (*i.e.*, dependent subproblems), the problem can not be properly decomposed [10, 11]. In this case, dependent subproblems may contain information from other subproblems. Consequently, a partition between them may interfere with the final result by adding a bias in the algorithm [7, 8].

Decision variables from spectroscopic data in multivariate calibration commonly present strong correlations among them [2]. As a consequence, our hypothesis arises:

- **Hypothesis**: Variable selection in multivariate calibration should not be treated as a decomposable problem due to the considerable data correlation (multicollinearity) usually present in the dataset.

### 3.1 Illustrative demonstration

In order to demonstrate the suitability of our hypothesis, we are providing two examples. The goal consists of demonstrating that Equation (4) is not satisfied in the context of variable selection in multivariate calibration.

**Example 1**. Let $S_{n\times4} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ be a subset with four random columns of matrix $\mathbf{X}_{n\times k}$:

$$S_{n\times4} = \begin{bmatrix} -0.0023 & 0.0013 & -0.0022 & -0.0013 \\ -0.0025 & 0.0014 & -0.0023 & -0.0014 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ -0.0020 & 0.0010 & -0.0020 & -0.0012 \end{bmatrix}.$$

Moreover, let $\mathbf{V}_{12\times4} = \{ \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4, \mathbf{v}_5, \mathbf{v}_6, \mathbf{v}_7, \mathbf{v}_8, \mathbf{v}_9, \mathbf{v}_{10}, \mathbf{v}_{11}, \mathbf{v}_{12} \}$ be a population with twelve individuals:

$$\mathbf{V}_{12\times4} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

where each individual selects different combinations of variables. Table 1 shows the separately-obtained RMSEP values by different combinations of variables in $S_{n\times4}$:

**Table 1: RMSEP values for different combinations of variables in matrix $S_{n\times4}$.**

| Variable subset | RMSEP |
|---|---|
| $\mathbf{v}_1 = \mathbf{x}_1$ | 10.7114 |
| $\mathbf{v}_2 = \mathbf{x}_2$ | 3.9159 |
| $\mathbf{v}_3 = \mathbf{x}_3$ | 6.1204 |
| $\mathbf{v}_4 = \mathbf{x}_4$ | 6.5862 |
| $\mathbf{v}_5 = \mathbf{x}_1 \cup \mathbf{x}_2$ | 3.9415 |
| $\mathbf{v}_6 = \mathbf{x}_1 \cup \mathbf{x}_3$ | 6.1242 |
| $\mathbf{v}_7 = \mathbf{x}_1 \cup \mathbf{x}_4$ | 6.5756 |
| $\mathbf{v}_8 = \mathbf{x}_2 \cup \mathbf{x}_3$ | 3.8765 |
| $\mathbf{v}_9 = \mathbf{x}_2 \cup \mathbf{x}_4$ | 3.9110 |
| $\mathbf{v}_{10} = \mathbf{x}_3 \cup \mathbf{x}_4$ | 5.8188 |
| $\mathbf{v}_{11} = \mathbf{x}_2 \cup \mathbf{x}_3 \cup \mathbf{x}_4$ | 3.4252 |
| $\mathbf{v}_{12} = \mathbf{x}_1 \cup \mathbf{x}_2 \cup \mathbf{x}_3 \cup \mathbf{x}_4$ | 3.4937 |

**Example 2**. Calculating the Pearson's linear correlation coefficient for matrix $S_{n\times4}$, one can obtain a symmetric matrix $\mathbf{R}_{4\times4}$ such as:

$$\mathbf{R}_{4\times4} = \begin{bmatrix} 1 & -0.3976 & 0.2498 & 0.2118 \\ -0.3976 & 1 & -0.0099 & -0.0218 \\ 0.2198 & -0.0099 & 1 & 0.9843 \\ 0.2118 & -0.0218 & 0.9843 & 1 \end{bmatrix}.$$

Pearson's linear correlation coefficient is usually represented by $\rho$ and takes values in the range [-1, 1]. Then, $\rho$ = 1 means a perfect positive correlation between two variables, $\rho$ = -1 means a perfect negative correlation, and $\rho$ = 0 means that both variables are not correlated [4].

These two examples indicate Equation (4) can not be satisfied for the variable selection problem in multivariate calibration. For instance, matrix $\mathbf{R}_{4\times4}$ points out that variables can influence each other due to the presence of multicollinearity among them. In matrix $\mathbf{R}_{4\times4}$, variables $\mathbf{x}_2$ and $\mathbf{x}_3$ provide $\rho$ = -0.0099 (close to zero), which means they are near linearly independent. Nevertheless, variables $\mathbf{x}_3$ and $\mathbf{x}_4$ have $\rho$ = 0.9843 (close to 1) and both are correlated. Hence, $\mathbf{x}_3$ and $\mathbf{x}_4$ are correlated to each other such that one variable may carry information from the other, which interferes with the condition that there can not be dependent subproblems.

We can notice the RMSEP values sum of each separate variable in Table 1 ($\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3 + \mathbf{x}_4$ = 27.3339) is considerably greater than the obtained RMSEP value with the four variables together ($\mathbf{x}_1 \cup \mathbf{x}_2 \cup \mathbf{x}_3 \cup \mathbf{x}_4$ = 3.4937). Then, it is possible to claim that $f(S_{n\times4}) \neq f_{\mathbf{v}^1_{1\times4}}(S^1_{n\times4}) + f_{\mathbf{v}^2_{1\times4}}(S^2_{n\times4}) + f_{\mathbf{v}^3_{1\times4}}(S^3_{n\times4}) + f_{\mathbf{v}^4_{1\times4}}(S^4_{n\times4})$.

Finally, the subset composed of variables $\mathbf{x}_2$, $\mathbf{x}_3$ and $\mathbf{x}_4$ in Table 1 provides the lowest RMSEP ($\mathbf{x}_2 \cup \mathbf{x}_3 \cup \mathbf{x}_4$ = 3.4252). This implies that as variables $\mathbf{x}_3$ and $\mathbf{x}_4$ are correlated, they should be together in the same subset. In addition, since variable $\mathbf{x}_2$ is the one which has the lowest correlation degree with both variables $\mathbf{x}_3$ and $\mathbf{x}_4$, these three variables are able to reduce the prediction error. As a consequence, they should remain together in the same subset.

Therefore, examples 1 and 2 provide significant evidences that selecting variables in multivariate calibration indeed should be considered (or treated) as a non-completely decomposable problem, which supports our hypothesis.

## 4 EXPERIMENTAL

Real dataset employed in this work consists of whole-wheat grain samples obtained from vegetal material from occidental Canadian producers. Standard data were determined at the Grain Research Laboratory as in works of Paula *et al.* [7–9]. The data set for the multivariate calibration study consists of 775 Near Infrared (NIR) spectra of whole-kernel wheat samples, which were used as shootout data in the 2008 International Diffuse Reflectance Conference.

Protein concentration in the analysed samples was chosen as the property of interest. Spectra were acquired by a spectrophotometer in range of 400-2500 nanometers (nm) with a resolution of 2 nm. Kennard and Stone algorithm was applied to the resulting spectra to divide the samples into three sets: calibration, validation and prediction with 389, 193 and 193 samples, respectively.

## 5 RESULTS AND DISCUSSION

Figure 1 plots the NIR absorbance spectra of wheat samples obtained by Canadian producers using a spectrophotometer [1]. The spectra in the chart come from the calibration set ($\mathbf{X}_{cal}$) which contains 389 samples, and each sample has 690 variables. It is possible to check the absorbance variations from different properties contained in the wheat samples. In general, these variations can cause wave mutual

[1]Spectrophotometer is a device used to measure the amount of light reflected or absorbed by a sample object.

disturbance (interference) implying in considerable rapprochement (dependency) among variables from the spectra [2, 3, 5]. Note that the variable index is related to the wavelength. This chart indicates that in most of wavelength regions there may be a relatively large number of correlated variables.

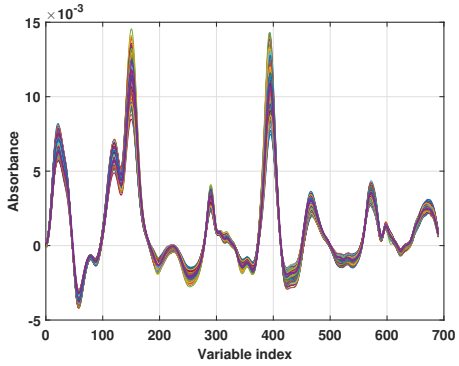


**Figure 1: NIR absorbance spectra of wheat samples.**

Figure 2 [2] shows a hot color map representing the correlations among all variables from Figure 1:

The *corrcoef Matlab*© built-in function yields a symmetric matrix $\mathbf{R}_{690\times690}$ calculated from the input matrix $\mathbf{X}_{389\times690}$ from Equation (2) whose rows are observations (samples) and columns are variables. This function is calculated by Pearson's linear correlation coefficient. The *imagesc Matlab*© built-in function displays the absolute value of elements from matrix $\mathbf{R}_{690\times690}$ as a symmetric image. In such image, the more elements close to 1 a variable vector has, the more correlated to other variables it is. Similarly, the closer to zero, the smaller the correlation degree.

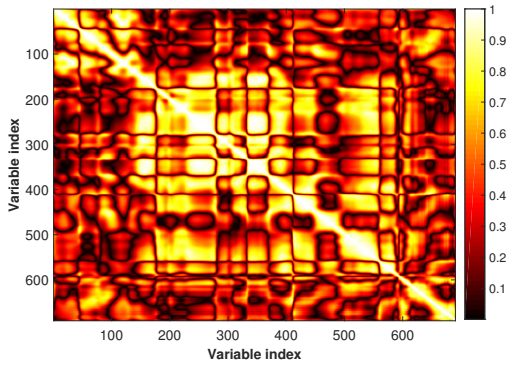


**Figure 2: Linear correlation analysis among all variables.**

One can notice in Figure 2 that in fact there are considerable correlations among most of variables. For example, variable 320 and variable 480 yield Pearson's linear correlation coefficient $\rho$ = 0.0172 indicating they are near linearly independent (orthogonal) and could be possibly selected to contribute for the increasing of model predictive ability. On the other hand, variables 250 and 350 yield $\rho$ = 0.9843 which indicates they are almost totally correlated and both contribute to increase the multicollinearity in the model. Hence, these variables should not be selected.

Finally, it is noteworthy that Figures 1 and 2 provide additional evidences about the non-decomposability assumption for the variable selection problem in multivariate calibration. Therefore, they strengthen our hypothesis.

[2]This chart was generated by using the *corrcoef* and *imagesc Matlab*© built-in functions.

## 6 WORK TO BE CONTINUED

Based on concepts of decomposability, a problem can be decomposed when it is split into independent subsets of variables. However, when one tries to split a subset of dependent variables, such problem may not be properly decomposed. Due to the constant presence of multicollinearity in spectroscopic data from multivariate calibration, our hypothesis claims that selecting variables in this context is a non-completely decomposable problem. Thus, our examples and outcomes provide significant evidences about the veracity of our hypothesis.

For the continuation of this work, we aim to develop a formal demonstration about our hypothesis. Additionally, we intend to improve our results previously published in [8] by proposing an enhanced local-based search GA in order to avoid the use of recombination operators. We have claimed that recombination operators used in standard GAs can cause building blocks disruption [8].

## ACKNOWLEDGMENTS

Authors thank to the Brazilian research agencies CAPES, FAPEG and CNPq for the financial support that has been provided.

## REFERENCES

[1] Chang Wook Ahn. 2006. *Advances in evolutionary algorithms.* Springer.
[2] Michael Patrick Allen. 2004. *The Problem of Multicollinearity In: Understanding Regression Analysis.* Springer Science and Business Media.
[3] Edward I. George. 2000. The variable selection problem. *J. Amer. Statist. Assoc.* 95, 452 (2000), 1304–1308.
[4] John Kalivas. 1993. *Mathematical analysis of spectral orthogonality.* Vol. 17. CRC Press.
[5] Harald Martens. 1991. *Multivariate calibration.* John Wiley & Sons.
[6] Ali Niazi and Riccardo Leardi. 2012. Genetic algorithms in chemometrics. *Journal of Chemometrics* 26, 6 (2012), 345–351.
[7] Lauro Cassio Martins Paula, Anderson Soares, Telma Lima, and Clarimar Coelho. 2016. Feature Selection using Genetic Algorithm: An Analysis of the Bias-Property for One-Point Crossover. In *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion.* ACM, 1461–1462.
[8] Lauro Cassio Martins Paula, Anderson Silva Soares, Telma W. Lima, Clarimar Jose Coelho, and Arlindo R. G. Filho. 2016. Variable Selection for Multivariate Calibration in Chemometrics: A Real-World Application with Building Blocks Disruption Problem. In *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference (GECCO).* ACM, 1031–1034.
[9] Lauro Cassio Martins Paula, Anderson Silva Soares, Telma W. Lima, Alexandre C. B. Delbem, Clarimar J. Coelho, and Arlindo R. G. Filho. 2014. A GPU-Based Implementation of the Firefly Algorithm for Variable Selection in Multivariate Calibration Problems. *Plos One* 9, 12 (2014), e114145.
[10] Martin Pelikan, David E Goldberg, and Erick Cantu-Paz. 2000. Linkage problem, distribution estimation, and Bayesian networks. *Evolutionary computation* 8, 3 (2000), 311–340.
[11] Franz Rothlauf. 2011. *Design of modern heuristics: principles and application.* Springer Science & Business Media.
[12] Richard A Watson, Gregory S Hornby, and Jordan B Pollack. 1998. Modeling building-block interdependency. In *Int. Conf. on Parallel Problem Solving from Nature.* Springer, 97–106.
[13] Huirong Xu and Bing Qi. 2012. Var. selec. in vis. and NIR spectra: Appl. to on-line determ. of sugar content in pears. *Journal of Food Engin.* 109, 1 (2012), 142–147.