UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

LAURO CÁSSIO MARTINS DE PAULA

Variable Selection in Multivariate Calibration considering Non-Decomposability Assumption and Building Blocks Hypothesis

Goiânia, GO - Brazil
2018

sibi sistema de bibliotecas ufg

PRPG
PRÓ-REITORIA DE PÓS-GRADUAÇÃO

UFG

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

1. Identificação do material bibliográfico: **[ ] Dissertação** **[x] Tese**

2. Identificação da Tese ou Dissertação:

Nome completo do autor: Lauro Cássio Martins de Paula

Título do trabalho: Variable Selection in Multivariate Calibration considering Non-Decomposability Assumption and Building Blocks Hypothesis

3. Informações de acesso ao documento:

Concorda com a liberação total do documento [x] SIM [ ] NÃO[1]

Havendo concordância com a disponibilização eletrônica, torna-se imprescindível o envio do(s) arquivo(s) em formato digital PDF da tese ou dissertação.

Lauro Cássio Martins de Paula
Assinatura do(a) autor(a)[2]

Ciente e de acordo:

Andrea S Evaristo
Assinatura do(a) orientador(a)[2]

Data: 06 / 12 / 2018

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. A extensão deste prazo suscita justificativa junto à coordenação do curso. Os dados do documento não serão disponibilizados durante o período de embargo.

Casos de embargo:
- Solicitação de registro de patente
- Submissão de artigo em revista científica
- Publicação como capítulo de livro
- Publicação da dissertação/tese em livro

[2] A assinatura deve ser escaneada.

LAURO CÁSSIO MARTINS DE PAULA

Variable Selection in Multivariate Calibration considering Non-Decomposability Assumption and Building Blocks Hypothesis

Thesis presented to the postgraduate program of Instituto de Informática from Universidade Federal de Goiás, as a partial fulfillment of the requirements for the Ph.D. degree in Computer Science.

Concentration area: Computer Science.

Advisor: Prof. Dr. Anderson da Silva Soares

Co-Advisor: Prof. Dr. Clarimar José Coelho

Goiânia, GO - Brazil
2018

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Cássio Martins de Paula, Lauro
Variable Selection in Multivariate Calibration considering Non Decomposability Assumption and Building Blocks Hypothesis [manuscrito] / Lauro Cássio Martins de Paula. - 2018.
CXVI, 116 f.

Orientador: Prof. Dr. Anderson da Silva Soares; co-orientador Dr. Clarimar José Coelho.
Tese (Doutorado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação em rede (UFG/UFMS), Goiânia, 2018.
Bibliografia.
Inclui siglas, abreviaturas, símbolos, gráfico, tabelas, algoritmos, lista de figuras, lista de tabelas.

1. Multivariate Calibration. 2. Variable Selection. 3. Genetic Algorithm. 4. Building Blocks. 5. Decomposability. I. da Silva Soares, Anderson, orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL
UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA
DOUTORADO EM CIÊNCIA DA COMPUTAÇÃO

UFMS

Ata de Defesa de Tese de Doutorado

Aos seis dias do mês de dezembro de dois mil e dezoito, no horário das dezenove horas, foi realizada, nas dependências do Instituto de Informática da UFG, a defesa pública da Tese de Doutorado do aluno Lauro Cássio Martins de Paula, matrícula no. 2015 0150 , intitulada **"Variable Selection in Multivariate Calibration considering Non-Decomposability Assumption and Building Blocks Hypothesis"**.

A Banca Examinadora, constituída pelos professores:

Prof. Dr. Anderson da Silva Soares – INF/UFG - orientador

Prof. Dr. Clarimar José Coelho – ECEC/PUC-GO - coorientador

Prof. Dr. Celso Gonçalves Camilo Júnior – INF/UFG

Prof. Dr. Fabrízzio Alphonsus Alves de Melo Nunes Soares – INF/UFG

Prof. Dr. Anselmo Elcana de Oliveira – IQ/UFG

Prof. Dr. Paulo Henrique Ribeiro Gabriel – FACOM/UFU

emitiu o resultado:

(X) Aprovado

( ) Aprovado com revisão

(A Banca Examinadora deve definir as exigências a serem cumpridas pelo aluno na revisão, ficando o orientador responsável pela verificação do cumprimento das mesmas.)

( ) Reprovado

com o seguinte parecer: _____

_____

_____

_____

Prof. Dr. Anderson da Silva Soares

Prof. Dr. Clarimar José Coelho

Prof. Dr. Celso Gonçalves Camilo Júnior

Prof. Dr. Fabrízzio Alphonsus A. M. N. Soares

Prof. Dr. Anselmo Elcana de Oliveira

Prof. Dr. Paulo Henrique Ribeiro Gabriel

Lauro Cássio Martins de Paula (bibliograhic citation: PAULA, L. C. M.)

Has a Bachelor's degree in Computer Science with emphasis in Computational Mathematics from Pontifical Catholic University of Goiás - Brazil, and a Master in Computer Science from Federal University of Goiás - Brazil. During his Doctorate, he received a CAPES scholarship and published important papers, which contributed to the development of this PhD thesis. Currently, he operates in the following research topics: scientific computing, parallel computing, evolutionary algorithms, bio-inspired metaheuristics for variable selection, Big Data and machine learning.

This PhD thesis is especially dedicated to our fabulous God, my amazing parents, my dear aunt and my deceased cousin Carlos Henrique da Silva (Cacau) who will always be in our hearts.

Esta tese de doutorado é especialmente dedicada ao nosso adorável Deus, meus maravilhosos pais, minha querida tia e meu falecido primo Carlos Henrique da Silva (Cacau), que estará para sempre em nossos corações.

Agradecimentos

I apologize to those that do not understand the brazilian portuguese language. I decided not to write this acknowledgment in english in respect to my family members who were always by my side during these years of study.

Para mim, chegar até aqui é a realização de um sonho pessoal, o qual surgiu há dez anos durante a minha graduação. Com isso, quero registrar os meus sinceros agradecimentos a todos aqueles que têm ajudado direta ou indiretamente na conquista de meus objetivos. Em primeiro lugar, agradeço a Deus por sempre me proporcionar saúde, paz de espírito e sabedoria para lutar todos os dias por um futuro cada vez melhor. A Ele, toda honra e toda glória!

Agradeço aos meus pais, Carlos Gardel de Paula e Beti Martins Borges, minha tia Dalva Aparecida Borges e meus falecidos avós Divina e Tomé por estarem sempre ao meu lado, independente da situação, fornecendo provas suficientes de amor incondicional. Gosto de ressaltar sempre que, sem eles, eu não seria quem eu tenho me tornado.

Agradeço ao Prof. Dr. Anderson da Silva Soares, meu orientador acadêmico desde o mestrado, pela boa pessoa que tem sido comigo. Por sua causa, tenho conseguido realizar muitos de meus objetivos acadêmicos e profissionais. Por exemplo, tive a oportunidade de apresentar artigos científicos em quatro países diferentes, além de ter despertado em mim um enorme interesse pela Ciência de Dados. Independente do tempo, jamais irei esquecer todo o esforço e toda confiança depositada em mim.

Agradeço ao meu co-orientador acadêmico, Prof. Dr. Clarimar José Coelho, por todos os seus ensinamentos desde a época da minha graduação. Participando do grupo de pesquisa em computação científica na Pontifícia Universidade Católica de Goiás (PUC Goiás), tive a oportunidade de aprender e crescer bastante. A sua paciência e colaboração foram fundamentais para o desenvolvimento deste trabalho.

Não poderia deixar de agradecer a minha noiva, Gisele Cardoso da Silva, pelo seu amor, carinho e compreensão. Te conhecer foi a melhor coisa que aconteceu na minha vida pessoal. Considero que Deus te colocou no meu caminho para me mostrar que sempre existe esperança e que o amor sempre prevalece. Eu te amo muito e irei continuar te fazendo uma mulher super feliz.

Como sempre, faço questão de lembrar e agradecer aos meus amigos Leonardo

Barra Santana de Souza e Leandro Barra Santana de Souza, os quais são os responsáveis por despertar em mim a vontade de me tornar um cientista. Se não fosse os tempos de inciação científica, muito provavelmente eu teria seguido outro caminho. Minha gratidão a vocês será eterna.

Agradeço aos meus amigos(as): Kelton de Sousa Santiago, Leonardo Afonso Amorim, Roussian Di Ramos Alves Gaioso, Marcella Scoczynski Ribeiro Martins, André Novaes, Marcos Vinicius Fernandes Calazans, Cleiton Luiz Correa de Sousa, e meu irmão Felipe Martins Correa de Sousa. Apesar do pouco tempo disponível que temos para interagir, vocês têm sido bons companheiros nessa longa jornada. Entre tantas idas e vindas, encontros e desencontros, acabamos sempre reservando um tempo para tomar um café ou conversar sobre diversos assuntos. Em especial, agradeço ao Kelton por ter despertado em mim uma renovação da mente e espiritual. Sem dúvidas, tudo isso me ajudou a ter tranquilidade e manter a paz interna para continuar firme na pesquisa.

Agradeço aos meus colegas da graduação na PUC Goiás: Lorena Alves dos Santos, Lino Barros, Roneidson José, Letícia Azevedo, Douglas de Freitas, André Luis, Roberto Mendonça, Francisco de Assis, Heber Nogueira, Rhelcris Salvino, Mirella Esther, Warllson Santos, Wanderson Arantes, Gilvan Vieira, Vinícius Santos, Vinícius Barcelos, dentre outros. Embora cada um tenha seguido seu próprio caminho, os momentos felizes que tivemos juntos serão eternamente lembrados.

Agradeço aos meus amigos de infância: Wander Scalia Gonçalves, Gustavo Camargo da Silveira, Álvaro Ribeiro e Átido Ribeiro. Apesar do nosso afastamento (com exceção do Wander), vocês marcaram a minha vida para sempre e jamais esquecerei de vocês. Espero um dia poder ter a oportunidade de reencontrá-los para conversarmos e matarmos a saudade.

Agradeço a todos os professores que tive até hoje. Com toda a certeza, a função que vocês desempenham na sociedade é fundamental para a vida de qualquer pessoa. Também agradeço aos professores membros da banca examinadora por todas as colaborações e considerações, as quais contribuíram para o aprimoramento deste trabalho. Em especial, agradeço ao Prof. Dr. Anselmo Elcana de Oliveira, do Instituto de Química da Universidade Federal de Goiás, por toda a sua ajuda nos meus momentos de dúvidas.

Um agradecimento especial ao povo brasileiro por ter financiado a minha pesquisa durante esses quatros anos. Ao contrário do que muitos acreditam, foram as altas taxas de impostos que todos nós pagamos que permitiram a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) me contemplar com uma bolsa de estudos, a qual me forneceu condições financeiras para realizar essa tese de doutorado.

Por fim, deixo o meu sincero agradecimento a todos aqueles que, infelizmente, não consegui lembrar de mencionar aqui, mas que, de alguma forma, têm contribuído para o meu constante crescimento intelectual, social, espiritual e pessoal.

“Science serves to give us an idea of how big is our ignorance.”

Felicité Robert de Lamennais,

.

Abstract

de Paula, Lauro Cássio Martins. **Variable Selection in Multivariate Calibration considering Non-Decomposability Assumption and Building Blocks Hypothesis**. Goiânia, GO - Brazil, 2018. 116p. PhD. Thesis . Instituto de Informática, Universidade Federal de Goiás.

The procedure used to select a subset of suitable features in a given data set consists in variable selection, which is important when the dataset contains large number of variables and many of them are redundant. Multivariate calibration combines variable selection with statistical techniques to build mathematical models which relate the data to a given property of interest in order to predict this property by selecting informative variables. In this context, variable selection techniques have been widely applied to the solution of several optimization problems. For instance, Genetic Algorithms (GAs) are easy to implement and consist in a population-based model that uses selection and recombination operators to generate new solutions. However, usually in multivariate calibration the dataset present a considerable correlation degree among variables and this provides an evidence about the problem not being properly decomposed. Moreover, some studies in literature have claimed genetic operators used by GAs can cause the building blocks (BBs) disruption of viable solutions. Therefore, this work aims to claim that selecting variables in multivariate calibration is a non-completely decomposable problem (hypothesis 1) as well as that recombination operators affects the non-decomposability assumption (hypothesis 2). Additionally, we are proposing two heuristics, one local search-based operator and two versions of an Epistasis-based Feature Selection Algorithm (EbFSA) to improve model prediction performance and avoid BBs disruption. Based on the performed inquiry and experimental results, we are able to endorse the viability of our hypotheses and demonstrate EbFSA can overcome some traditional algorithms.

Keywords

Multivariate Calibration, Variable Selection, Genetic Algorithm, Building Blocks, Decomposability.

Resumo

de Paula, Lauro Cássio Martins. **Variable Selection in Multivariate Calibration considering Non-Decomposability Assumption and Building Blocks Hypothesis**. Goiânia, GO - Brazil, 2018. 116p. Tese de Doutorado . Instituto de Informática, Universidade Federal de Goiás.

Seleção de variáveis é um procedimento para selecionar um subconjunto de características viáveis em um conjunto de dados, o qual se torna importante quando esse conjunto contém muitas variáveis redundantes. A calibração multivariada combina seleção de variáveis com técnicas estatísticas para construir modelos matemáticos com o intuito de predizer uma propriedade de interesse. Nesse contexto, técnicas de seleção têm sido aplicadas na solução de diversos problemas. Por exemplo, Algoritmos Genéticos (AGs) são fáceis de implementar e consistem em um modelo baseado em população, o qual utiliza operadores de seleção e recombinação para gerar novos indivíduos. No entanto, geralmente em calibração multivariada, o conjunto de dados apresenta um grau de correlação considerável entre as variáveis e isso nos fornece uma evidência de que tal problema não pode ser decomposto adequadamente. Além disso, alguns estudos da literatura têm afirmado que os operadores genéticos utilizados pelos AGs podem causar o rompimento dos Blocos Construtores (*Building Blocks* - BBs) das soluções viáveis. Portanto, este trabalho objetiva demonstrar que a seleção de variáveis em calibração multivariada é um problema não-completamente decomponível (hipótese 1), assim como que operadores de recombinação afetam a presunção de não-decomponibilidade (hipótese 2). Adicionalmente, este trabalho propõe duas heurísticas, um operador de busca local e duas versões de um Algoritmo para Seleção de Variáveis baseado em Epistasia (EbFSA) para aprimorar a capacidade de predição do modelo e evitar o rompimento de BBs. Baseando-se na pesquisa realizada e nos resultados obtidos, torna-se possível confirmar a viabilidade de nossas hipóteses e demonstrar que o EbFSA consegue superar alguns algoritmos tradicionais.

Palavras–chave

Calibração Multivariada, Seleção de Variáveis, Algoritmo Genético, Building Blocks, Decomponibilidade.

Contents

List of Figures

List of Tables

List of Algorithms

List of Symbols

$\mathbf{X}$	Matrix of observations and variables	19
$\mathbf{y}$	Vector of reference values	19
β	Vector of regression coefficients	25
ε	Portion of random error	25
n	Number of observations	26
q	Radiation absorbance frequency	27
P_0	Radiation emitted by a spectrophotometer	27
P	Radiation absorbed by sample	27
λ	Wavelength	27
k	Number of columns in matrix $\mathbf{X}$	25
ρ	Pearson's linear correlation coefficient	38

List of Abbreviations and Acronyms

SPA	Successive projections algorithm	32
NIR	Near-Infrared	20
BB	Building Block	21
GA	Genetic algorithm	34
CGA	Compact Genetic algorithm	49
MLR	Multiple linear regression	19
PLS	Partial Least Squares	32
PLSR	Partial Least Squares Regression	19
PCR	Principal Component Regression	19
RMSEP	Root Mean Squared Error of Prediction	25
LL	Linkage Learning	50
EbFSA	Epistasis-based Feature Selection Algorithm	73

CHAPTER 1

Introduction

Variable selection is the procedure used to choose a subset of suitable features contained in a given dataset. Selecting variables becomes important when the dataset contains many redundant and irrelevant features, which do not provide distinguished knowledge and should be removed without incurring loss of information [39]. For example, in machine learning and statistics variable selection is commonly used for constructing a model from which it is possible to be interpreted by users. Machine learning is an area of study from artificial intelligence. It aims to develop algorithms and techniques which allow the computer to learn some task. For some applications, the learning can be defined in specific terms. In this context, the multivariate calibration arises in machine learning which uses statistical techniques to build mathematical models that establish learning about data [66].

Multivariate calibration is a sub-area of study from chemometrics [1] related to analytical chemistry. It determines a mathematical model which relates the data to a given property of interest from a known sample set in order to predict this property by selecting informative variables [66]. Given a set of explanatory variables in a matrix $\mathbf{X}_{n\times k} = \mathbf{x}^1_{n\times 1}, \mathbf{x}^2_{n\times 1}, \ldots, \mathbf{x}^k_{n\times 1}$, with n observations and k variables, and given a set of response variables in a vector $\mathbf{y}_{n\times 1} = (y_1, y_2, \ldots, y_n)^T$, the learning task [2] can be synthesized to find $\mathbf{y} = f(\mathbf{X})$. Among the main statistical techniques used to perform the calibration process and obtain mathematical models are: partial least squares regression (PLSR) [111], principal component regression (PCR) [38], neural networks (NN) [20], locally weighted regression (LWR) [21], radial basis function combined with PLS (RBS-PLS) [114], and multiple linear regression (MLR) [70]. MLR [3] is a statistical technique used to build models which describe the relationships among several informative variables [23, 72]. It works in the data original domain and is particularly important for some types

[1] Basically, chemometrics is an analytical chemistry field that applies statistical and mathematical methods to problems from chemistry. It uses mathematical tools to design or select experimental procedures.

[2] In the context of multivariate calibration, $\mathbf{y}$ is called vector of reference values and $\mathbf{X}$ is called matrix of variables and observations, where each observation is a row and each variable is a column of such matrix.

[3] It is important to emphasize that in the context of this work, only MLR is to be utilized.

of applications which require an understanding about the produced learning model. Although PLSR and PCR have been applied in multivariate calibration, MLR is widely used and can be a good alternative to turn data into information [23].

In order to turn data into information, it often becomes needed to perform variable selection. The variable selection procedure arises in the construction of mathematical models to establish the relationship between $\mathbf{y}_{n\times 1}$ and a subset of variables from $\mathbf{X}_{n\times k}$ when one is not sure about which subset is to be utilized. This issue is particularly interesting when the number of columns in $\mathbf{X}_{n\times k}$ is large and $\mathbf{X}_{n\times k}$ contains many redundant variables [25]. Then, variable selection may be considered as a statistical application problem. Such problem commonly recurs in multivariate calibration. Moreover, variable selection is an important procedure in chemometrics problems involving regression and classification tasks. For classification, variable selection allows a better separation among categories and assists in the construction of a mathematical model able to describe the different categories with good specificity and sensitivity [25]. In a regression problem, one can build a mathematical model able to explain a response variable [3]. Thus, the use of variable selection techniques becomes a viable alternative to settle such issue.

To deal with larger and more complex datasets, the development of efficient variable selection methods becomes an increasingly important asset. Such methods aim to search a combination of variables to produce the best result [4] by eliminating variables which produce noise or that, although giving good information, are strictly correlated with other already selected variables. In this context, several studies have proposed algorithms for the variable selection procedure in multivariate calibration. For instance, Xu *et al.* [121] presented a genetic algorithm (GA) implementation for variable selection in visible (VIS) and near-infrared (NIR) spectra. Four different variable selection methods were used to determine the sugar content of pears. Authors showed that the proposed GA can be used for industrial applications.

Arakawa *et al.* [5] proposed a GA-based wavelength selection method for spectral calibration, where the goal consisted on the construction of robust and predictive regression models by selecting informative wavelength regions. They showed that their proposal works better than other traditional techniques. Nevertheless, although the proposed algorithm can be used in combination with any regression method, it was used only partial least squares (PLS) method in such study.

Niazi and Leardi [74] 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 reference on the subject. In their paper, it is possible to see the applications of GAs in three main different areas: *i*) optimization; *ii*) structure-activity

[4] For example, explaining the property of interest from the instrumental data (*e.g.*, an analyzed sample).

relationship (QSAR) and molecular modeling; and *iii*) multivariate calibration.

On the other hand, some studies have claimed genetic operators used by recombination-based search methods such as GA usually cause the disruption of the building blocks (BBs) during their executions [41, 47, 62, 88, 91]. BB is a low-order and short defining-length schema with an above-average fitness representing a set of solutions to a subproblem [41]. Disrupting BBs often leads the set of solutions from GAs to be trapped in local optima besides increasing the computational time [47]. For example, some authors have demonstrated crossover operators used by standard GAs tend to cause the building blocks disruption, which usually leads to undesirable performance [13, 47, 62, 84].

1.1 Work Proposal

On the one hand, it is known the selection of a non-correlated (independent) variable subset implies in the improvement of the model predictive ability. On the other hand, in practice this does not always work well because the number of variables is usually large and the correlation among them is commonly strong [18, 23, 31]. For instance, the value of only one independent variable may be considerably worse than the value obtained from a subset of independent variables due to the variance of the variables [116]. Furthermore, even if one tries to select only the best variables (*e.g.*, the most informative), the BBs disruption caused by genetic operators in recombination-based search algorithms may affect the conjecture of non-decomposability when the problem is not decomposable [96].

1.1.1 First hypothesis

Decomposable problems can be treated by concatenating basis functions of a certain order [2]. To a problem to be decomposable, there must be none interaction between any two variables and each variable should be separately treated [94]. According to Watson [115], included in the term decomposable is the notion of identifiable component parts, and a set of correlated variables is not decomposable. Often in multivariate calibration there are considerable linear dependency among decision variables from spectral data [6, 66]. Then, based on such statements it becomes possible to realize the variable selection in multivariate calibration usually can not be properly decomposed.

In this sense, our first hypothesis arises and claims that spectral data in multivariate calibration may be considered as a non-completely decomposable problem. In a non-completely decomposable problem, some variables are closely correlated to other variables and therefore should be maintained in the same subset [115]. In this case, variable selection procedure in multivariate calibration may not be accordingly treated as a

decomposable problem due to the constant presence of multicollinearity in the dataset (see Section 2.2). Therefore, based on a comprehensive bibliographic review about concepts of decomposability, Section 4.1.2 presents Hypothesis 1 and Equation (4-2) together with three numerical examples in order to point its viability.

1.1.2 Second hypothesis

In general, schemata are not considered as a critical for variable selection because their use depends on an a priori knowledge about patterns to be pursued in the schemata formation [88, 107]. If schemata are properly set, the convergence tends to be faster. Otherwise, it can limit the search space and result in loss of performance [41]. The BBs hypothesis appeals to the notion of problem decomposition and the assembly of solutions from sub-solutions. The method of forming solutions by first breaking down a problem into subproblems is a central tenet behind the BBs hypothesis [115].

In this context, our second hypothesis arises. It claims that not necessarily there exists BBs formation in spectral data from multivariate calibration due to the high data dimensionality. However, the schemata disruption caused by recombination operators in standard GAs can directly affect the non-decomposability assumption of the variable selection procedure in multivariate calibration (which is raised up by Hypothesis 1). Therefore, based on a deep research about schema theory, Section 4.2.1 presents Hypothesis 2 and uses one proposition together with three additional numerical examples aiming to show its interference in the first hypothesis.

It is important to note the numerical examples for both hypotheses lack some theorem in order to demonstrate their generalization. Such issue extrapolate the scope of this work (see Section 7.2). The goal consists in showing their feasibility and applicability into our two large datasets (see Chapter 5).

1.1.3 Proposed implementations

We are providing significant outcomes using a proposed GA implementation to show additional evidences about the non-decomposability assumption and schemata disruption problem in multivariate calibration models. In the proposed GA implementation (Section 4.3), we are applying two heuristic strategies and one simple local search operator. Such approaches aim to improve the GA implementation through empiricism [5].

The first heuristic (Section 4.4.1) consists in the initial individuals generation, which selects the best chromosomes (schemata candidates) as initial solutions. This

[5]The strategies applied into the both heuristics and the local search operator are performed on an empirical manner.

simple strategy is based on the Successive Projections Algorithm (see Section 2.3.2) and is achieved by using projection operations and selecting (near) orthogonal variables.

The second heuristic aims to identify possible schemata. It is a naive strategy and tries to find the schema which possibly generates the best individuals in the population. Considering the best schema as the one capable of generating $x\%$ of the best individuals, where x can be empirically chosen, it becomes possible to calculate the schema fitness (see Section 3.1).

The local search operator is an approach based on the Variable Neighborhood Search [94]. It replaces the crossover operator and performs a simple local search by modifying particular genes in the chromosomes. Genes are bits in our binary representation, and each bit is equivalent to a variable. The local search explores the variable neighborhood and reduces (but not eliminate) the schemata disruption.

Additionally, we are proposing two versions of a novel approach for variable selection. It is called Epistasis-based Feature Selection Algorithm (EbFSA). EbFSA (Section 4.6) uses two different enhanced strategies. The concept of epistasis (Section 3.5.3) becomes important to measure and analyze the genes interdependence. To this end, we are using the Pearson's linear correlation coefficient (Section 2.4.3) as the epistatic relation among the variables in our both datasets (dataset 1 and dataset 2). Consequently, EbFSA applies a deterministic approach and avoids the schemata disruption by keeping correlated variables together in the same subset. Avoiding the schemata disruption is crucial to prevent the problem decomposition, especially when dealing with high data dimensionality and interdependent subproblems (which is our case).

Based on our broad inquiry and significant experimental results, we are able to evidence the viability of both hypotheses. Moreover, our proposed EbFSA is able to select the most informative variables, providing the best outcomes and overcoming traditional algorithms in terms of variable selection as well as model predictive ability. Finally, all results obtained in this work are scientifically published or in peer review process (see Section 7.1).

1.2 Work Organization

The remainder of this work is organized as follows. Chapter 2 introduces the mains concepts about variable selection in multivariate calibration. The BB hypothesis as well as schema theory are discussed in detail in Chapter 3. Our proposal is presented in Chapter 4. Chapter 5 provides the materials and methods used to obtain the experimental results. All outcomes are presented and discussed in Chapter 6. Finally, Chapter 7 presents our conclusions, contributions, limitations and suggestions for future work.

CHAPTER 2

Variable Selection in Chemometrics

Chemometrics is the science of extracting information from chemical systems by data-driven means [14]. In general, it refers to the application of statistical methods into problems from chemistry [9]. Chemometrics is commonly applied to solve both descriptive and predictive problems. In descriptive applications, properties of chemical systems are modeled with the intent of learning the underlying relationships and structure of the system. In predictive applications, properties of chemical samples are modeled with the intent of predicting new properties or behavior of interest [66]. In both cases, datasets are often large and highly complex involving hundreds or thousands of features. Feature (or variable) selection is the automatic selection of attributes in datasets which are most relevant to the predictive modelling problem.

Section 2.1 describes the main concepts regarding multivariate calibration. Section 2.2 depicts all details about an inconvenience called multicollinearity. Some techniques for variable selection such as Genetic Algorithm (which is used in this work) are described in Section 2.3. Finally, Section 2.4 introduces some multicollinearity assessment procedures [1].

2.1 Multivariate Calibration

One of the sub-areas of study from chemometrics related to analytical chemistry is the multivariate calibration. Multivariate calibration is the procedure used to determine a mathematical model able to predict some properties of interest from known samples by an instrument such as a spectrophotometer. It can be represented by Figure 2.1.

Initially, the sample is inserted in the spectrophotometer. Spectral lines (*e.g.*, near-infrared spectroscopy) are released onto the sample. Through the absorption or reflection process of the spectra by the sample molecules, it becomes possible to obtain

[1] One of them is the Pearson's linear correlation coefficient, which is chosen to be used in this work.

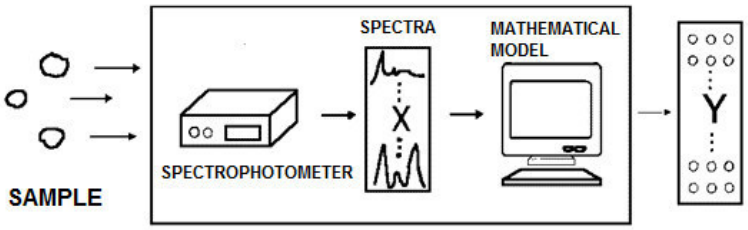

Figure 2.1: *Representation of multivariate calibration process, adapted from [33].*

data about this relationship. Then, a mathematical model can be obtained to measure the concentration level of a property of interest from the sample [2].

Such a mathematical model establishes the relationship between the properties measured by the spectrophotometer and the concentration of an analyzed sample [66]. It can be used to provide the value of a quantity $\mathbf{y}$ based on values measured from a set of explanatory variables $\mathbf{X}_{n\times k} = \{\mathbf{x}^1_{n\times 1}, \mathbf{x}^2_{n\times 1}, \ldots, \mathbf{x}^k_{n\times 1}\}$, and can be defined by Equation (2-1):

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \ldots + \beta_k \mathbf{x}_k + \varepsilon, \tag{2-1}$$

where β_0, β_1, ..., β_k are the regression coefficients to be determined, and ε is a measure of random error.

In order to obtain the coefficients in Equation (2-1), one may use the multiple linear regression (MLR). MLR is a statistical technique used to build models which describe the relationships among several explanatory variables [23, 72]. Equation (2-2) shows how those regression coefficients can be calculated using the Moore-Penrose pseudoinverse [60]:

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

where $\mathbf{X}_{cal}$ [3] is the $n \times k$ matrix of variables and observations from the calibration set, $\mathbf{y}$ is the vector of reference variables, and β is the vector of regression coefficients.

Soon after calculating those coefficients in Equation (2-2), it becomes necessary to determine the predictive ability of MLR models [4]. 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

[2] For instance, the level of protein in wheat samples could be considered as the property of interest.

[3] Kennard and Stone [57] algorithm can be applied to divide matrix $\mathbf{X}_{n\times k}$ into three sets: calibration ($\mathbf{X}_{cal}$), validation ($\mathbf{X}_{val}$) and prediction ($\mathbf{X}_{pred}$). See Chapter 5 for more details.

[4] Usually, reference values (previously yielded in laboratory) can be used to assess the model predictive ability.

predicted by a model and the values actually observed [110]. In the context of multivariate calibration, it is depicted such as shown in Equation (2-3):

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

where $\mathbf{y} = \{y_1, y_2, ..., y_k\}^T$ is the reference values of the property of interest (which is attempted to be determined in the analyzed sample), n is the number of observations (number of rows of matrix $\mathbf{X}_{cal}$), and $\hat{\mathbf{y}} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_k\}^T$ is the estimated value calculated in matrix notation as

$$\hat{\mathbf{y}} = \mathbf{X}_{val}\beta, \tag{2-4}$$

where $\mathbf{X}_{val}$ is the $n \times k$ matrix of variables and observations from the validation set.

It is also possible to use other criteria such as the Mean Absolute Percentage Error (MAPE) and the Predicted Residual Sums of Squares (PRESS) in order to determine the predictive ability of MLR models. MAPE is one relative measure describing erros as a percentage of the actual data. It can be used to measure how high or low are the differences between predictions and actual data in regression models [85]. This measure is defined by Equation (2-5):

$$MAPE = \frac{\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right|}{n}(100), \tag{2-5}$$

where y_i is the actual data at variable i, $\hat{y}_i$ is the forecast (using some model/method) at variable i, and n is the number of observations (or samples).

PRESS is an useful statistical measure used for comparison between different models. It also can be used as a predictivity measure to compare and select the best model. Equation (2-6) shows how to calculate it:

$$PRESS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{2-6}$$

where each y_i is the real value of concentration obtained by laboratorial methods, $\hat{y}_i$ is the result of Equation (2-4) applied to measures of new observations ($\mathbf{X}_{cal}$), and n is the number of observations.

Based on the error prediction value [5], it becomes possible to determine if the model has or not an adequate predictive ability. Generally, the goal consists in obtaining a model with a considerably reduced error value. However, in order to achieve this goal it

[5] There are several statistical measures available in literature. However, in this work we are using RMSEP as the main statistical measure. We do not intend to compare model prediction error between different measures.

often becomes necessary to deal with the multicollinearity problem. For instance, the light absorbance by the sample at a given wavelength can be related to the examined compound concentration. Figure 2.2 shows the absorption process assessing the radiation absorbance with frequency q in a spectrophotometer. The absorption intensity is given by Equation (2-7).

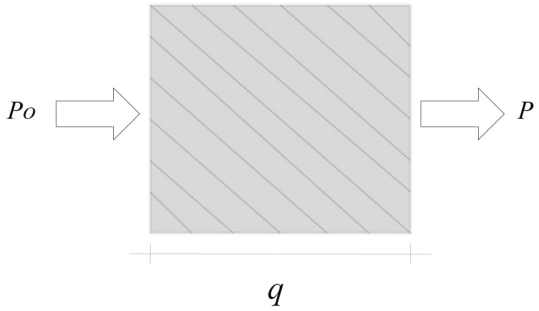

Figure 2.2: *Absorption spectroscopy process, adapted from [75].*

$$A(\lambda) = log\frac{P_0(\lambda)}{P(\lambda)}, \tag{2-7}$$

where $P_0(\lambda)$ is the emitted light intensity by the device, and $P(\lambda)$ is the absorbed light intensity by the sample at wavelength λ.

When two or more wavelengths cause interference in each other, a mutual disturbance in waves may occur. The overlap of waves usually causes a high correlation (linear dependency) among variables [6], which can imply in mathematical problems such as multicollinearity in the calibration-model achievement process [39]. In addition, a large number of variables can be obtained for a given sample causing matrix $\mathbf{X}_{n\times k}$ to have a greater number of columns than the number of rows and resulting in an ill-conditioning of the matrix. An ill-conditioned matrix can cause divergence problems and directly affect the prediction quality of the property of interest from the sample being analyzed [18].

2.2 Multicollinearity

One of the main issues related to the calibration process is the recurrent presence of linear correlation among variables. The existence of linear correlation between two or more variables is a mathematical problem defined as multicollinearity [23]. Multicollinearity can be caused by the relationship among explanatory variables, and it is an undesirable attribute of the particular calibration set being collected. Multicollinearity

[6] In this work, variables are sampled waves, *i.e.*, each wavelength represents a specific variable.

can reduce the reliability of coefficients from estimated models [3]. Moreover, the reliability reduction of coefficients usually implies in an ill conditioning of matrix $\mathbf{X}_{n\times k}$ [23]. A poorly conditioned matrix can present invertibility problems and result in an inverse matrix which does not match the expected actual values.

2.2.1 Spectral orthogonality

Spectroscopic data can contain multicollinearity for several reasons. In this sense, two variable vectors are said to be collinear if they both lie on a line passing through an origin [10]. The concept of orthogonality serves to quantify how close a given vector is to being dependent upon a set of vectors [31]. For example, consider two vectors **a** and **b**. Thus, their dot product is calculated as

$$\mathbf{a}\cdot\mathbf{b} = cos\theta||\mathbf{a}||_2||\mathbf{b}||_2, \tag{2-8}$$

where $0 \leq \theta \leq 180^\circ$ is the angle between **a** and **b**, and $||\mathbf{a}||_2||\mathbf{b}||_2$ is the euclidian norm.

Vectors **a** and **b** are orthogonal if $\theta = 90^\circ$ [10]. The value of an angle provides quantitative information about the collinearity (dependency) between two vectors [56]. By definition, a set of vectors $\{\mathbf{x}^1_{n\times 1}, \mathbf{x}^2_{n\times 1}, \ldots, \mathbf{x}^k_{n\times 1}\}$ is defined to be dependent (correlated) if at least one of the vectors $\mathbf{x}_i$, $1 \leq i \leq k$, can be written as linear combination of the remaining vectors. Similarly, if those vectors are not correlated, they are linearly independent.

For instance, let $\{\mathbf{x}^1_{n\times 1}, \mathbf{x}^2_{n\times 1}, \ldots, \mathbf{x}^k_{n\times 1}\}$ be the columns of the $n \times k$ matrix $\mathbf{X}_{cal}$ in Equation (2-2). In addition, let $\{\sigma_1, \sigma_2, \ldots, \sigma_p\}$, $1 \leq p \leq min\{n,k\}$, be the singular values [7] of $\mathbf{X}_{cal}$. If $p = k$, the columns of $\mathbf{X}_{cal}$ are said to be linearly independent, meaning $\mathbf{X}_{cal}$ is a full rank matrix [56]. The rank of a matrix is the dimension of the vector space generated by its columns. Thus, the column rank of $\mathbf{X}_{cal}$ is the dimension of its column space. According to Bourbaki [10], matrix $\mathbf{X}_{cal} \in R^{n\times k}$ is a full column rank if and only if $\mathbf{X}^T_{cal}\mathbf{X}_{cal}$ is invertible.

2.2.2 Sources of multicollinearity

There are five main multicollinearity sources [3, 18, 31, 56]:

- model specification;
- calibration design deficiencies;
- overdefined models;

[7] In mathematics, the singular values are the square roots of the eigenvalues, which are scalars associated with the eigenvector [56].

- outlier-induced multicollinearity;
- multicollinearity inherent in the sample.

Model specification multicollinearity usually results from variable definitions. It exists regardless of the sample being analyzed [56]. For example, adding polynomial terms to a model can cause linear dependency [70]. Variable selection techniques can be used to redefine the model [18]. Such approach is commonly used by biased regression methods such as principal components and partial least squares (see Section 2.3).

Calibration design multicollinearity is an attribute of the particular calibration set being collected [23]. It is not inherent in the model or sample. If it has multicollinearity because calibration samples lie in a proper subspace of the intended predictive domain, it is said to be deficient [56]. According to Kalivas [56], the best way to treat it is to augment the dataset. This is because collection of additional calibration samples can improve spectral orthogonality in matrix $\mathbf{X}_{cal}$, which usually implies in the reduction of linear correlation [31]. However, the worst situation for this procedure consists on removing unnecessary variables, and the deletion of such variables can generate a bias in calibration matrices and predictions [18, 56].

An overdefined model usually has more variables than samples in the calibration process, *i.e.*, $k > n$. It is an undesirable situation because it can produce multicollinearity [3]. In the context of multivariate calibration, it means there are more regression coefficients than samples which can be used to estimate them. This can occur with spectroscopic data because current spectrophotometers can supply absorbance measurements for a sample at hundreds or thousands of wavelengths in a few seconds [87]. Multicollinearity can be smoothed if more calibration samples are used for matrix $\mathbf{X}_{cal}$ during calibration process. However, adding more calibration samples will not always provide a feasible solution, and more variables may need to be selected [6, 48].

Outliers can cause artificial multicollinearity. It can derive from calibration design deficiencies [56]. Multicollinearity can be induced if an outlier implies large values on two or more variables. The removal of outliers reduces multicollinearity. Nevertheless, it is necessary to identify a real outlier before removing it from dataset [48]. Otherwise, removing false outliers may imply in the loss of significant information.

Sample-inherent multicollinearity comes from sample characteristics and usually can not be avoided by using any sampling procedure or experimental design [56]. Even if the calibration samples are well designed, they may not necessarily yield an orthogonal set. One of the first steps in a proper treatment of the multicollinearity problem consists on its detection or diagnosis [23]. An alternative to minimize this type of near dependency consists on performing wavelength (or variable) selection. Variable selection can reduce the degree of multicollinearity [39]. For instance, many multivariate calibration models present poor performance when multicollinearity exists among variables [85, 101, 121,

124]. Thus, it becomes necessary to deal with multicollinearity by using some technique able to select the most informative and non-redundant variables. The removal of non-informative variables produces better predictions and simpler models. According to Xiaobo [119], a well-performed variable selection can result in models having a greater predictive ability.

2.2.3 Dealing with multicollinearity

In literature, it is possible to find many techniques to deal with multicollinearity [3, 18, 23, 31, 119]. According to Kalivas [56], sometimes the best thing to treat multicollinearity is doing nothing. Indeed, this is far away from being the best solution for handling multicollinear data. Notwithstanding, if an analysis sample has much multicollinear data as the calibration sample, it is expected the concentration prediction of the property of interest should be reasonably accurate [3]. Such approach is clearly limited in its predictive ability, and it demands to be carefully decided.

As mentioned before, augmenting the data may smooth the multicollinearity. However, augmenting multicollinear datasets is not a practical choice since multicollinearity usually stems from the sample [56]. Moreover, additional data may only yield the same dependency, besides collecting new data is not always feasible [31]. A possible alternative to minimize multicollinearity consists on the respecification of the model by combining variables. Respecifying the model by variable combination may reduce the degree of dependence among them [56]. For instance, consider $\{\mathbf{x}^1_{n\times 1}, \mathbf{x}^2_{n\times 1}, \mathbf{x}^3_{n\times 1}\}$ as the first three columns of matrix $\mathbf{X}_{cal}$ from Equation (2-2). If $\{\mathbf{x}^1_{n\times 1}, \mathbf{x}^2_{n\times 1}, \mathbf{x}^3_{n\times 1}\}$ are linearly dependent (correlated), then a new variable $\mathbf{x}^*_{n\times 1} = \mathbf{x}^1_{n\times 1} + \mathbf{x}^2_{n\times 1} + \mathbf{x}^3_{n\times 1}$ could be used as a function to retain the information in the three variables in order to reduce the dependency. In this case, since these three variables are linearly dependent, they should be removed from the model because they tend to increase multicollinearity. Consequently, they have minimum predictive value. Instead, rather than eliminating them, the combination of such variables into one variable may yield a more predictive one [48].

If a decision has been made to perform variable selection, it becomes necessary to decide on which variables to utilize. In this sense, several variable selection algorithms are available in literature (see Section 2.3). To decide on which selection approach is the best depends on the selection criterion and the dataset complexity [56]. For example, in multivariate calibration, variable selection attempts to identify and remove correlated variables which reduce the model performance. Variable selection procedures are important when dealing with spectroscopic data [102]. In spectroscopic data, the number of variables is often large and needs to be reduced [37, 85, 101]. When MLR is being used, variable selection is an important step for building the calibration model.

Usually, methods such as Partial Least Squares Regression (PLSR) [111] is used in combination with MLR to reduce the dimensionality problem. Nevertheless, PLSR latent variables can be affected by the presence of irrelevant variables [116]. Thus, the use of enhanced techniques can improve the prediction performance. Finally, understanding the chemical process under investigation can be facilitated by paying attention to the relevant spectroscopic variables [56, 119]. As a consequence, improvements in prediction performance can be expected when variable selection is being applied.

Another significant issue related to multivariate calibration consists in the fact that, in predicting problems with regression models containing many variables, most of them often may not contribute for the predictive ability of the model [124]. Thus, selecting a reduced set of informative variables becomes important to improve the efficiency of techniques used to construct MLR models. However, some independent variables are more significant than others [66]. Better outcomes are usually improved by eliminating uninformative variables. Furthermore, lower-cost predictions can be achieved by reducing the number of independent variables. Hence, smaller RMSEP values can be achieved.

In this sense and as mentioned before, the use of variable selection methods has become important approaches to deal with multicollinearity [6, 18, 85, 101]. The goal of variable selection consists in improving the prediction performance of predictors as well as providing a better understanding of the underlying process which generated the data. To address this issue, several techniques for variable selection have been proposed [5, 6, 85, 101, 121].

An approach consists on the use of filter and wrapper methods [48]. Filter method selects subsets of variables as a pre-processing step independently of the chosen predictor. Wrapper method uses the learning machine of interest as a black box to score variable subsets according to their predictive ability. Finally, a more recent approach consists in the use of adapted metaheuristics for variable selection. Some works in the literature have demonstrated their effectiveness and superiority over traditional variable selection methods [64, 77, 82, 85, 103, 101]. Section 2.3 provides a more detailed discussion about this issue.

2.3 Some Techniques for Variable Selection

There are many types of variable selection techniques. The choice of variable selection approach commonly depends on the problem. On the one hand, if data analysis needs to be performed in high dimensional space, selecting filter approach seems to be a better alternative to avoid high computational costs [48]. Examples of some filter methods include the Chi-squared test [90], information gain [58], correlation coefficient scores [71], and novel feature selection approaches [24, 109, 112].

On the other hand, if accuracy is more desirable, wrappers should be used [59]. One of the advantages of wrapper methods is the estimated accuracy of the learning algorithm is the best available heuristic for measuring the variable values [24]. Examples of wrapper methods are the recursive-feature elimination algorithm [48], partial least squares [111], successive projections algorithm [6], among others [20].

2.3.1 Partial least squares

Partial least squares [111], also know as PLS, is a technique which generalizes and combines variables from principal component analysis and MLR. PLS can yield a more (statistically) reduced and robust solution than MLR [49, 95]. When a relatively large number of correlated variables are present in the model, PLS is useful for reducing noisy predictor variables [111]. It provides a set of informative scores about the correlation structures of the variables and structural similarities among compounds, which means that it bears some relation to principal components regression [95].

Instead of finding hyperplanes of minimum variance in correlated and independent variables, it is able to find a linear regression model by projecting the predicted and observable variables to a new space [111]. In general, PLS creates and uses score vectors. Such vectors are called latent variables and are defined to be maximized by the covariance among different sets of variables. However, one of its main problems lies in the fact that it guides the variables transformation by covariance in which the relationship of variance and covariance does not necessarily lead to better learning for some applications [32, 116, 124]. Furthermore, in case of considerable large number of correlated variables, PLS can cause overfitting obtaining a well-fitted model with a reduced predictive ability [6].

Commonly, PLS is considered as a biased regression method [32, 56]. Although consisting in a strategy which attempts to reduce multicollinearity by removing uninformative variables, it usually skews the final results and tends to introduce a bias in the model by eliminating some of the original information [56]. Then, a strict test for the significance of each PLS component becomes necessary [95]. According to Kalivas [56], it is hoped that when using a biased regression method, estimations will be closer to actual values than estimations based on unbiased regression. Nevertheless, in practice this does not always occur. In literature, one can find works that have used other approaches with PLS [5, 32, 124].

2.3.2 Successive projections algorithm

Successive projections algorithm (SPA) is an iterative procedure used for variable selection [6]. It is a forward selection method that uses mathematical operations in

a vector space to minimize multicollinearity. In the context of multivariate calibration, SPA is used for variable selection in MLR models (SPA-MLR) [34, 87, 105]. In its general form, SPA-MLR is composed of three phases. In the first phase, the instrumental responses of the calibration samples are disposed in matrix **X** from Equation (2-2). Each column of **X** is considered as a different variable (column vector). Such column vectors are subjected to a sequence of projection operations. Each projection operation results in the creation of chains of variables, and each element of a chain is included so as to obtain the largest orthogonal projection [79].

In phase 2, the candidate subsets of variables extracted from the chain created in phase 1 are evaluated. The best subset of variables is selected based on the smallest prediction error value by using some statistical measure (*e.g.*, RMSEP from Equation (2-3)). Finally, the last phase involves a backward elimination procedure which discards those (uninformative) variables that do not present significant information [105]. To this end, a relevance index is defined for each variable included in the subset selected in the previous phase. Such index is calculated by multiplying the standard deviation of the variable by the absolute value of its regression coefficient [6]. Then, variables are sorted according to the index, and a chart plotting a comparison between RMSEP and the number of variables can be generated.

Several works have used SPA for selecting features from optimization problems. For instance, Araujo [6] proposed the SPA for variable selection in spectroscopic component analysis. The results demonstrated that in comparison with PLS, SPA was able to provide better outcomes and a higher computational performance. By using SPA, Breitkreitz [12] presented a strategy for sulfur determination in diesel samples. Author showed that in comparison with a genetic algorithm, SPA yielded better MLR models with a more adequate prediction ability.

Due to the need of constructing a MLR model for each candidate subset of variables, phase 2 of SPA can be considered as the computational bottleneck of the algorithm [79]. Moreover, the matrix inverse calculation in Equation (2-2) can demand a significant computational effort. Consequently, the computational performance can be reduced. Thus, some works have presented strategies to minimize the computational time of phase 2 [79, 87, 104]. In this sense, Paula *et al.* [79] proposed a new strategy implemented in SPA to avoid matrix inverse calculation in order to increase the computational performance. Results presented 37$\times$ better performance compared to a traditional SPA implementation.

On the other hand, other studies have claimed SPA is not able to overcome adapted metaheuristics for variable selection [64, 77, 82, 85, 86, 101]. Such works have demonstrated the use of metaheuristics such as Genetic Algorithm (see Section 2.3.3) can provide better outcomes regarding the prediction error reduction in multivariate

calibration models as well as reducing the number of selected variables.

2.3.3 Genetic algorithm

Despite deterministic approaches have been widely used for variable selection, some studies have demonstrated the superiority of metaheuristics over such methods [64, 85, 101]. Possible advantages over deterministic methods are: in many cases, the search space is considerably large; and most metaheuristic strategies provide a search space exploration in parallel with multiple solutions. Hence, the computational performance tends to increase when a metaheuristic is being used [122].

Genetic algorithm (GA) is an evolutionary metaheuristic based on recombination process in which each individual represents a specific solution. It consists in an optimization approach based on Darwin's classical rules about natural evolution. Such approach uses random steps in order to converge to a nonrandom optimal solution [2, 41]. GA may be considered as a population-based model by using selection and recombination operators to generate new sample points in a search space. Some algorithms also use mutation operators, which allow the candidate solutions to not be tied in local optima [42].

An implementation of a GA begins with a population of random candidate solutions. Such solutions are also called chromosomes. Each chromosome is evaluated based on a fitness function, and those which represent a better solution to the target problem are given more chances to reproduce than those one which are poorer solutions. Figure 2.3 shows an example of two chromosomes randomly chosen to participate in the recombination process using an one-point crossover operator. In such figure, both subject 1 and subject 2 were generated by a binary representation in which "1" may indicate that a variable is selected, and "0" otherwise. Algorithm 2.1 shows one simple pseudocode for GA.

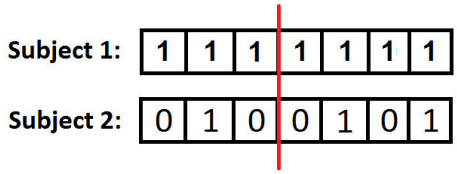

Figure 2.3: *Example of two chromosomes in a recombination process using a single-point crossover.*

Recombination process results in the next generation of chromosomes which are (usually) different from the previous generation. Generally, the average fitness is increased by this procedure since only the best chromosomes from the first generation are selected for breeding along with a small proportion of less fit solutions. These less fit solutions ensure genetic diversity within the genetic pool of the parents, and therefore ensure the genetic diversity of the subsequent offspring [41].

Algorithm 2.1: Simple genetic algorithm.

```
Let n be the population size
Initialize the population of individuals using uniformly distributed random numbers
for i = 1 : MaxGenerations
    Evaluate all individuals based on the fitness
    Generate n solutions using some crossover operator
    if offspring fitness is better than parents fitness
        Replace parents by children
    else
        Ignore children
    end if
end for i
```

Some works in literature have used GA for variable selection to solve different types of optimization problems. For instance, Arakawa *et al.* [5] proposed a GA-based wavelength selection method for spectral calibration. The goal consisted in the construction of robust and predictive regression models by selecting informative variables. It was shown that their proposal works better than other traditional techniques. Niazi and Leardi [74] published a review which covers the application of GAs in chemometrics. The objective was showing the main research fields of GAs applications, together with providing a list of reference on the subject. In their paper, it is possible to see the applications of GAs in different areas. More recently, Soares *et al.* [101] proposed a mutation-based compact genetic algorithm (mCGA) for spectroscopy variable selection. Authors demonstrated that mCGA is able to minimize the prediction error and requires a reduced number of evaluations than a standard GA.

In general, GAs are easy to implement. GAs have the ability to produce even more fitter partial solutions by combining building blocks (BBs) [41]. However, their behavior is difficult to understand due to its stochasticity, and recombination operators used by standard GA implementations tend to cause BBs disruption over generations [47, 84] (see Chapter 3). As a consequence, significant information may be lost through crossing operations among individuals of the population. Furthermore, it is difficult to understand why these algorithms frequently succeed at generating solutions of high fitness when applied to practical problems [108]. In this context, it is important to highlight that this work uses the GA as a tool to investigate and analyze such inconveniences in multivariate calibration (see Chapter 4).

2.4 Multicollinearity Assessment

It is known that multicollinearity causes serious issues in the calibration process as well as in the model predictive ability. Hence, it becomes useful to use a diagnostic

method to detect and assess the presence of multicollinearity in the dataset [56]. Furthermore, the use of variable selection techniques helps to discern on which variables are linearly dependent (*i.e.*, involved in multicollinearity), and select those who provide better information for the property of interest to be accurately and precisely estimated from the analyzed sample (see Section 2.3).

Euclidian distance among spectra is usually applied as a match indicator to determine spectral similarity [113]. In this sense, the distance measurement between two variables (spectral vectors) may be a good choice to assess the collinearity among spectra. However, assessments based on distance measurements are not always worthwhile methods [113]. As described in Section 2.2.1, spectral orthogonality can be assessed in order to appraise multicollinearity. Dot product is an example of tool which can be used to assess orthogonality between two vectors. Unfortunately, only the use of dot product to assess orthogonality may be a naive procedure for being considered limited in its usefulness [56, 113].

Distance measurements and dot products may not be practical choices for assessing spectral orthogonality. Kalivas [56] claims that there are some assessment procedures which represent the most useful evaluators of multicollinearity. Some procedures are functions of singular values from a singular value decomposition of the matrix under consideration (matrix $\mathbf{X}_{cal}$ in case of multivariate calibration). Using this type of methodology, it becomes possible to assess the degree of multicollinearity in the dataset. Three examples of measures commonly utilized are: *i*) condition number; *ii*) absolute determinant; and *iii*) Pearson's linear correlation coefficient.

2.4.1 Condition number

Condition number of a matrix can be a valuable asset to evaluate multicollinearity. It has the ability to correlate data with error in concentration predictions. Equation (2-9) shows how the condition number of matrix $\mathbf{X}_{cal}$ can be computed:

$$\kappa_2(\mathbf{X}_{cal}) = \frac{\sigma_1}{\sigma_k}, \tag{2-9}$$

where σ_1 and σ_k are, respectively, the first and the last singular values of matrix $\mathbf{X}_{cal}$.

It is possible to notice if $\kappa_2(\mathbf{X}_{cal})$ is large relative to σ_1, then σ_k is small. The term σ_k represents the distance from $\mathbf{X}_{cal}$ to all $n \times k$ matrices with dependent columns [56]. Then, if $\kappa_2(\mathbf{X}_{cal})$ is large relative to σ_1, $\mathbf{X}_{cal}$ is close to a $n \times k$ matrix whose columns are dependent. Instead, the smaller the $\kappa_2(\mathbf{X}_{cal})$ value, the greater the independence between any two columns of $\mathbf{X}_{cal}$.

However, Kalivas [56] experimentally demonstrated the potential inability of condition numbers to assess spectral orthogonality. In many cases, it is the size of

the singular value σ_k that represents a more critical assessment. Moreover, singular values lack the ability to estimate the extent of spectral orthogonality for a specific component (column) in $\mathbf{X}_{cal}$ and can not point exactly which columns are involved in multicollinearity [56].

2.4.2 Absolute determinant

Determinant of a matrix is a function associating a scalar with each square matrix. This function lets us know if the matrix has or not an inverse matrix. If its determinant is equal to zero, then such matrix is singular [10]. For example, one can prove there is a single function f with the following properties: *i*) f is linear and alterning in the rows of the matrix; and *ii*) $f(\mathbf{I}_n) = 1$, where $\mathbf{I}_n$ is the identity matrix. Such function f is called determinant and the determinant of matrix $\mathbf{X}_{cal}$ can be represented by $det(\mathbf{X}_{cal})$ [10].

On the other hand, the absolute determinant is defined to be the singular values product of the matrix under consideration. For instance, consider $\mathbf{X}_{cal}$ as a full rank $n \times k$ matrix ($n \geq k$) with singular values $\{\sigma_1, \sigma_2, \ldots, \sigma_k\}$. If one or more dependent variables are present in the matrix, one or more of the singular values tend to be small (close to zero) [10, 56]. The absolute determinant is defined for a $n \times k$ matrix with full rank k such as

$$Det(\mathbf{X}_{cal}) = \prod_{i=1}^{k} \sigma_i. \tag{2-10}$$

The determinant and absolute determinant of a matrix are distinct concepts. The determinant of a matrix is defined if and only if the matrix is square [10]. The absolute determinant is defined for $n \times k$ matrices with full rank k [56]. Through Equation (2-10), it is possible to see that an absolute determinant is always positive since all $\sigma_i > 0$, $1 \leq i \leq k$. Instead, the determinant of a matrix can be zero, positive or negative. If matrix $\mathbf{X}_{cal}$ is non-singular (*i.e.*, has an inverse matrix), then the absolute determinant and the determinant of $\mathbf{X}_{cal}$ can be calculated by Equation (2-11):

$$Det(\mathbf{X}_{cal}) = det(\mathbf{X}_{cal}). \tag{2-11}$$

From the absolute determinant, it follows that a near zero absolute determinant value reflects near dependency among columns of $\mathbf{X}_{cal}$ [56]. Similar to condition number, although having the potential for assessing dependent variables, the absolute determinant also can not determine exactly which columns are involved in multicollinearity [3]. Consequently, it should be considered as a global assessment tool instead of a fine analysis tool [56].

2.4.3 Pearson's linear correlation coefficient

Another relevant measure used to assess multicollinearity is the Pearson's linear correlation coefficient (see Example 3 in Section 4.1.3). The Pearson's linear correlation coefficient is a statistical measure for the linear correlation between two variables [97]. It can yield a symmetric matrix $\mathbf{R}_{k\times k}$ from the $n \times k$ matrix $\mathbf{X}_{cal}$ with values -1 $\leq \rho \leq$ 1, where ρ = -1 represents a total negative linear correlation, ρ = 0 indicates no linear correlation between two variables, and ρ = 1 is a total positive linear correlation. The ρ value of two variables $\mathbf{x}_1$ and $\mathbf{x}_2$ ($\rho_{\mathbf{x}_1,\mathbf{x}_2}$) is their covariance divided by the product of their standard deviations [63]:

$$\rho_{\mathbf{x}_1,\mathbf{x}_2} = \frac{E[(\mathbf{x}_1 - \mu_{\mathbf{x}_1})(\mathbf{x}_2 - \mu_{\mathbf{x}_2})]}{\omega_{\mathbf{x}_1}\omega_{\mathbf{x}_2}}, \tag{2-12}$$

where μ is the mean, ω is the standard deviation, and E is the expectation.

In addition to being simple and relatively easy to calculate, the Pearson's linear correlation coefficient is not a naive measure such as condition number. Moreover, it does not demand considerable computational effort [63]. Therefore, it is important to highlight that in this work we use this measure to assess and analyze the linear dependence among variables (columns) in matrix $\mathbf{X}_{cal}$ (see Sections 4.6 and 6.3 for more details).

2.5 Summary

This chapter discussed the main concepts about the variable selection procedure in multivariate calibration. Multivariate calibration is a field of study in chemometrics used to assess properties of interest from samples by mathematical models. It is possible due to the use of instrumental devices (*e.g.*, spectrophotometer), which emits infrared spectra over the analyzed sample. Through these absorbed or reflected spectra by the molecules in the analyzed sample, a mathematical model can be established by using statistical techniques such as multiple linear regression. However, current spectrophotometers are able to generate thousands of variables and many of them do not present significant information about the property of interest. This issue is mainly caused by a mathematical problem called multicollinearity. It can arise from five main different sources: inherent in the sample, model specification, calibration design deficiencies, overdefined models, and outlier-induced multicollinearity.

Multicollinearity causes linear correlation among variables. This issue reduces the reliability of estimated models. Consequently, it must be treated in order to improve the model predictive ability. Spectral orthogonality measurement between two variables may be a viable tool for quantifying how close a given variable is to being dependent upon a set of variables. Condition number, absolute determinant and the Pearson's linear

correlation coefficient can be valuable assets to assess and evaluate the collinearity between two or more variables.

The use of variable selection techniques has been a more suitable choice to deal with multicollinearity. Genetic Algorithm is an example of metaheuristic approach which has been widely used to select variables in optimization problems. In this context, Chapter 2 aimed to introduce the basic concepts about variable selection in multivariate calibration in order to help the reader understand our proposal (which is described in Chapter 4).

CHAPTER 3

Building Blocks Hypothesis

The term building block (BB) has been used without a formal definition, but it was initially cited by Holland [53] and Goldberg [47]. BB is defined as a short and low-order schema [1] with an above average fitness. More specifically, a BB can be described as a subset of solutions to a subproblem expressed as a schema. Such schema has high fitness and its size is smaller than the length of the binary solution. BBs can be helpful for estimating the performance of recombination-based search methods such as genetic algorithms [94]. The building blocks hypothesis can be formulated by using the definition of schemata as being highly-fit solutions to subproblems. According to Goldberg [41], short, low-order and highly-fit schemata should be sampled and recombined to form solutions of potentially higher fitness.

In order to grasp the BB problem, next sections detail the main concepts regarding such issue. Section 3.1 provides the theory behind the concept of schema. Section 3.2 introduces the schema theorem proposed by Goldberg [41]. An introduction about the generalization of the schema theorem is presented in Section 3.3. Section 3.4 presents a discussion about the disruption of building blocks, establishing relation with problem difficulty, deceptive functions, disruption caused by crossover operators, and the possible formation of variables blocks in spectral data. Finally, Section 3.5 argues about some alternatives such as competent genetic algorithms, linkage learning and epistasis, which are possible solutions for avoiding BBs disruption.

3.1 Schema Theory

A schema $H = (h_1, h_2, ..., h_l)$ is a string of length l and may be considered as a template representing a subset of solutions. If a solution is a string built using symbols from an alphabet, one can say schemata are strings made by such alphabet plus the character $*$ (don't care). The $*$ character represents any given character from the existing

[1] Schema is a genetic pattern describing a set of chromosomes in the search space with certain fixed positions [53].

alphabet. For example, $H = 1******$ can be considered as a schema for the subject 1 from Figure 2.3 (see Section 2.3.3). In this case, schema H is able to yield any string in which the first position is fixed to 1 and the other positions do not care.

In schema theory, there are two important characteristics which must be taken into consideration: *i*) the order $O(H)$ of a schema H is defined as the number of fixed positions in the string; and *ii*) the defining length $\delta(H)$ of a schema H is the distance between the two outermost fixed positions [41]. For instance, consider H_1 and H_2 as the following schemata in a binary alphabet. In this case, H_1 has $O(H_1) = 3$ and $\delta(H_1) = 4$. Schema H_2 has $O(H_2) = 2$ and $\delta(H_2) = 3$. Strings 11011 and 01100 can be considered as instances of schemata H_1 and H_2, respectively.

$$H_1 = 1*0*1,$$
$$H_2 = *1**0.$$

The fitness of a schema H consists in the average fitness of all instances matching the schema [42]. It can be calculated as

$$f(H) = \frac{1}{|H|} \sum f(\mathbf{s}), \forall\, \mathbf{s} \in H, \tag{3-1}$$

where $|H|$ is the number of individuals $\mathbf{s}$ which are instances of schema H, and f is the fitness function.

The number of individuals that can be generated by a schema H is calculated as

$$m(H) = 2^{l-O(H)}, \tag{3-2}$$

where l is the string length [41].

The framework of schema theory allows the definition of a building block and a formal statement of the building block hypothesis [69, 107]. In schema theory, the search space is partitioned into subspaces of varying generality levels, and mathematical models are constructed to estimate how the number of individuals in the population belonging to certain schema can be expected to grow in the next generation. From these models, the BB hypothesis arises [69]. BB hypothesis attempts to explain how a GA solves a problem by positing that near-optimal solutions are forged from small, low-order and fitter-average schemata [69, 107, 118]. Based on the notion of BB hypothesis, Holland [53] developed the schema theorem which permits the design of GAs and the choice of their parameters. Such theorem describes how the number of instances of a schema H changes over the number of generations.

3.2 Schema Theorem

Schema theorem is commonly used to explain how GAs work. Typically, schema theorem states the expected proportion of individuals in a given schema at the current generation [91]. Under a particular genetic operation mechanism (proportionate selection) and a particular genetic operator (crossover operator), the schema theorem can be written as follows [53]

$$m(H,t+1) \geq m(H,t)\frac{f(H,t)}{f'(t)}\left[1-p_c\frac{\delta(H)}{l-1}\right], \tag{3-3}$$

where

- $m(H,t)$ is the expected number of individuals representing schema H at generation t;
- $f(H,t)$ is the fitness of schema H at generation t;
- $f'(t)$ is the average fitness of the population at generation t;
- p_c is the crossover probability (an algorithm parameter);
- $\delta(H)$ is the defining length of schema H;
- l is the string length.

Overall, schema theorem describes how the number of copies of a schema H depends on proportionate selection and crossover operator when using a standard GA [93]. It consists in the product of the selection ($m(H,t)\frac{f(H,t)}{f'(t)}$) and crossover ($[1-p_c\frac{\delta(H)}{l-1}]$) terms [41]. However, the usefulness of the schema theorem has been widely criticised [4, 19, 35, 91]. It provides only lower bounds for the expected value of the number of instances of a schema in the next generation [4]. This makes it difficult to predict the future behavior of a GA even for a single generation ahead [91].

Altenberg [4] points out schema theorem is not able to address the search component of GAs on which performance depends on. Moreover, it can not distinguish GAs that are performing well from those that are not [47]. In addition, Chung [19] states an important type of above-average schemata which are not BBs always exists among the schemata receiving increasing evaluations during the GA search, and the role of these schemata can not be explained by schema theorem.

In general, schema theorem represents two conflicting objectives: selection and crossover [41]. If the fitness of a schema H is above the average fitness of the population ($f(H,t) > f^{'}(t)$), selection tends to favor such schema. It happens because selection preserves highly-fit schemata to the next generations [54]. However, the defining length ($\delta(H)$) of a schema H should be small when using crossover [93]. Otherwise, crossover operator may frequently disrupt high-order or large-$\delta(H)$ schemata, which is a concern [47, 94, 108]. On the other hand, if the sub-solutions (schemata) to a problem are

short (low $\delta(H)$) and low-order (low $O(H)$), the number of desirable schemata tends to increase over generations [93]. Consequently, the problem may be easily solved by GA when schemata have high-average fitness and relatively-low crossover disruption.

Due to the schema theorem be restricted to a proportionate selection and a crossover operator, it may limit its own value [41]. Radcliffe [92] claims it is not possible to always explain the observed behavior of GAs. Schema theorem neglects the stochastic and dynamic nature of the genetic search [93]. Thus, Goldberg [41] introduced a generalization for this theorem. Basically, author substituted appropriate terms in order to ease the demonstration of the applicability through other selection mechanisms and genetic operators.

3.3 Generalization of the Schema Theorem

Goldberg [41] recognized that other selection methods may allocate the schemata on a different manner from proportionate selection. Then, a generalized selection pressure $\phi(H,t)$ is taken as a function of the reproduction ratio of schema H at generation t, and a disruption factor $\varepsilon(H,t)$ of schema H at generation t is put in place of $p_c\frac{\delta(H)}{l-1}$ as shown below:

$$m(H,t+1) \geq m(H,t)\frac{f(H,t)}{f'(t)}\phi(H,t)\left[1-\varepsilon(H,t)\right]. \tag{3-4}$$

The rightmost term $[1-p_c\frac{\delta(H)}{l-1}]$ in Equation (3-3) is a bound on the survival probability of a schema H under the applied genetic operators [47]. In the generalization form, this issue is mitgated by the use of $\phi(H,t)$ and $\varepsilon(H,t)$. However, the main concern still consists in making sure desirable schemata grow [41, 93]. This requires the growth factor γ is greater than one ($\gamma > 1$) [47, 68], where

$$\gamma = \phi(H,t)\left[1-\varepsilon(H,t)\right]. \tag{3-5}$$

In this sense, there are two possibilities:

- raizing the selection pressure $\phi(H,t)$; or
- decreasing the disruption factor $\varepsilon(H,t)$.

On the one hand, if $\phi(H,t)$ is high, an uniform population will be quickly obtained [93]. Thus, it will lead to the loss of diversity [68]. On the other hand, a reduced $\varepsilon(H,t)$ will decrease the probability of crossover, and this will lower the number of schemata exchange [94]. Therefore, in order to grow desirable schemata, one must combine viable schemata from one solution with viable schemata from another solution

using crossover, where the goal consists in minimizing the schemata disruption and maximizing the schemata exchange [107].

Old and recent works have questioned the appropriateness of the schema theorem and its generalization [13, 29, 30, 41, 47, 62, 65, 68, 93, 108]. One problem about the schema theorem is: due to fitness-proportional selection, schemata are not uniformly sampled by the population members [65]. A second problem about the schema theorem consists in the theoretical considerations derived from the evolution strategies do not confirm Holland's view on how genetic algorithms process building blocks [47]. Finally, a third problem is: the schema theorem formula can be interpreted in a way which seems to contradict the notion of the building blocks hypothesis [68, 93].

3.4 The Building Blocks Problem

BBs are helpful for estimating the performance of recombination-based search algorithms [94]. Routhlauf [94] claims that if the sub-solutions to a problem are short (low $\delta(H)$) and low order (low $O(H)$), the problem is assumed to be easy for recombination-based search. On the other hand, there are two main issues preventing the schema theorem to provide an exact analysis [117]. First, it is usually calculated an upper bound on the probability of disruption due to crossover. The term $\frac{\delta(H)}{l-1}$ assumes a schema is disrupted every time crossover operator falls within its defining length [41]. Second, the schema theorem ignores all sources of schemata from crosses of strings containing different competing schemata [13]. Furthermore, ensuring the growth factor $\gamma > 1$ becomes necessary to increasing the selection pressure or decreasing the disruption factor, which it is not always possible to achieve both accordingly.

There are other issues implying in the performance reduction of a recombination-based search algorithm. For example:

- Deceptive functions are commonly considered as difficult problems to be solved by GAs (see Section 3.4.1).
- High-quality but high-order schemata tend to be disrupted by crossover operators. As a consequence, they can not be inherit by next generations (see Section 3.4.2).
- In the context of spectral data in multivariate calibration, spectral regions representing the property of interest in the analyzed sample may be set in blocks and such blocks are usually broken due to schemata disruption caused by recombination operators [84] (see Section 3.4.3).

3.4.1 BB-based problem difficulty and deceptive functions

Previous work have shown some problems become more difficult to solve when using GA specific representations [68, 93, 94]. To be able to investigate how the representation influence the GA performance, it is necessary to apply a measurement of problem difficulty [41]. Focusing on recombination-based search GAs, it is possible to determine the reasons of problem difficulty consist in the building blocks [93].

Deceptive problems were one of the first approaches to the question of what makes problems difficult for GAs [47]. Indeed, GAs usually have difficulties in solving deceptive (or trap) problems [13, 62]. This occurs because individuals may be attracted to the trap solution since the fitness of all the solutions are nearest to the local optima instead of the global optimum [2]. Figure 3.1 shows the representation of an example of a 5-bit trap function. Considering the value 5 in the figure as the global optimum, the value 4 would be the deceptive attractor.

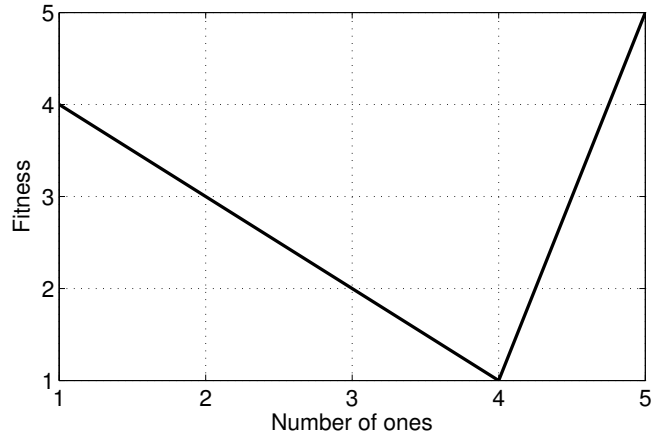

Figure 3.1: *Representation of a 5-bit trap function, adapted from [41].*

Deception is only one among several features of a problem affecting GA's performance [94]. Based on the schema theorem and the BB hypothesis, Goldberg [41] presented an approach to understand problem difficulty. This approach consists in the matter of schemata disruption, and it is decomposed into three parts [3]:

- intra-BB difficulty;
- inter-BB difficulty; and
- extra-BB difficulty.

[2] To reliably solve deceptive problems, GAs must increase the number of copies of the best schemata [93].

[3] Such problem-difficult decomposition assumes difficult problems are building block challenging [93].

The first one (intra-BB difficulty) means difficulty within a BB. A number of schemata of order $O(H) = k$ having the same fixed positions consists in 2^k different competing schemata. The different schemata compete against each other based on their fitness [41]. Identifying the high-quality schemata and properly propagating them is the main difficulty of intra-BB difficulty, and it can be measured with the deceptiveness of a problem [44].

The inter-BB difficulty means difficulty between two BBs. If a problem can be decomposed into smaller subproblems, GAs can solve these smaller subproblems in parallel and try to identify the desirable schemata [94]. However, the contributions of different schemata to the fitness function are not uniform, and interdependencies among different schemata usually occur [41]. Therefore, one of the major problems related to inter-BB difficulty is the loss of desirable schemata during a GA run. It happens due to the fact that different schemata can have different contributions to the fitness of an individual. Moreover, if a problem can not be decomposed into independent subproblems, there still can be interdependencies among different schemata, which is an additional issue of inter-BB difficulty [93].

Finally, extra-BB difficulty means difficulty outside of BBs. Noise may be considered as a source of extra-BB difficulty, and it has an important influence on the GA performance due to selection is based on the fitness of the individuals. Non-deterministic noise is often undesirable by modifying the fitness values of the individuals. Thus, selection decisions may be no longer based on the quality of the solutions. Instead, they may become based on stochastic variance [4][94].

3.4.2 Disruption due to crossover

Although the schema theorem and its generalization form are not self-sufficient to provide an accurate analysis of a GA, the principle of meaningful BBs is directly motivated by them [45]. Recall that if schemata are highly-fit, short and low-order, they tend to exponentially increase in numbers or market share over the generations [41]. However, if the high-quality schemata are long or high-order, they are often disrupted by crossover operators [93]. Hence, such schemata can not be properly propagated by recombination-based search methods.

For instance, consider two strings S_1 and S_2 both generated by schema $H =$ 01*01:

$$S_1 = 01101,$$

[4]Stochastic variance is a model in which the variance of a stochastic process is itself randomly distributed [123].

$$S_2 = 01001.$$

In this case, it is possible to observe that every possible cross of a crossover operator produces a copy of S_1 and S_2 as offspring. Thus, a recombination operator which always cross over a single differing bit returns a copy of the original strings between the two strings produced by the cross [47]. In this sense, high-order schemata are usually minimally disruptive. Nevertheless, it is not a useful operator because of the loss of diversity [29, 47, 108].

On the other hand, strings with two or more different bits (low-order schemata) usually cause the schema disruption for at least one cross site [47]. For instance, consider two random strings S_3 and S_4 generated by schema $H_1 = 1***0$ and schema $H_2 = 0***1$, respectively:

$$S_3 = 11100,$$
$$S_4 = 00011.$$

As claimed by Mitchell *et al.* [69], GAs work well when highly-fit and low-order schemata recombine to form even more highly-fit and low-order schemata. In this case, however, both schemata (H_1 and H_2) will be disrupted regardless where the cross site falls among the bits. Then, although smoothing the loss of diversity in the search space, such issue can move the solution away from the global optimum and get it stuck in local optima.

Schemata analysis is an important approach used for measuring the difficulty and disruption of subproblems with respect to GAs [46, 62]. As the main search operator of GAs is recombination, they are a representative example of recombination-based search techniques [93]. Schemata are commonly defined for binary search spaces and schemata analysis is useful for problems with binary representation and decision variables [94]. However, as far as we know, literature lacks studies describing how schemata concepts can be used for estimating problem difficulty as well as schemata disruption in the context of variable selection for spectral data.

3.4.3 Possible variables blocks in spectral data

Spectrophotometry is a procedure of emission and absorption of (*e.g.*, infrared) wavelength measurement by a device (*e.g.*, spectrophotometer). Using a spectrophotometer, it is able to determine the wavelengths or frequency of bands in a spectrum. The most widely used spectrometric methods make use of eletromagnetic radiation, which is a type of energy taking different forms. The visible spectrum (VIS) is one of the most easily recognized forms. Spectrophotometry has been used in applications such as the identity confirmation of a compound and the quantification of a protein [6, 85, 104]. The near

infrared (NIR) spectrophotometry offers a relatively-quick chemical analysis method capable of providing results of multiple properties in known samples.

Spectrophotometric relation between the absorbed energy by a chemical component and the variation of the wavelength stemmed by a spectrophotometer can be represented by the Lambert-Beer law [100]. As it is already known, the samples can contain different molecules with different sizes and complexities. When the spectrophotometer emits a particular wavelength onto the sample, a set of molecules resonates with the emitted frequency and absorbs (or reflects) such energy. Thus, by emitting multiple wavelengths it is possible to detect different types of absorptions (or reflections). As a consequence, it becomes possible to evaluate different sample properties.

The set of such emissions can generate a spectrum of the sample with respect to the emitted frequencies. In many cases, it is possible to check the absorbance variations from different properties (*e.g.*, protein) contained in the analyzed sample (*e.g.*, wheat grain). Nevertheless, it is not possible to obtain the concentration of a specific property of interest only with such spectra. In this scenario, multivariate calibration techniques are widely used to build mathematical models able to relate the spectra with the real concentration of the property of interest from the analyzed sample (see Chapter 2).

As described in Chapter 2, usually in multivariate calibration problems some independent variables are more significant than others [23]. Accuracy is improved by eliminating dependent variables, and lower-cost predictions can be achieved by reducing the number of independent variables. Hence, smaller prediction errors (*e.g.*, RMSEP) and a multivariate dataset can be more parsimoniously obtained [66]. Thus, it is known that the best variables are chosen according to their best performance in RMSEP (or other error measure), which is relevant given the fact that calibration model relates the spectrophotometric samples to reference samples. However, it is not possible to claim these variables represent the behavior of the actual concentration contained in the original sample. Such questioning may be relevant due to the possible information not included in the calibration model.

Even with techniques to reduce collinearity among variables (reducing instrumental noise, baseline errors, among others), it is still possible the model has information not referring to the original sample information. An example would be the existence of different molecules responding to the same stimuli (wavelengths) [100]. Then, if a variable contained in this spectral region is chosen, it could improve the calibration model (*e.g.*, reducing the prediction error), but with low concentration representation in the sample.

In this sense, we believe spectral regions reliably representing the properties of interest in the samples may be possibly set in blocks [84]. These blocks of variables can be interpreted as the ideal setting to represent the concentration of the property of interest

in samples. Therefore, blocks with the best variables (*e.g.*, those with best fitness) should be preserved during the model building process. Notwithstanding, schemata disruption caused by crossover operators becomes a critical issue to be taken into consideration [41, 47].

Finally, as far as we know, literature also lacks studies demonstrating the possible variables block formation and BBs setting in spectral data from multivariate calibration problems. It is important to note such issue is far away from the scope of this work (see Section 7.3). The main goal consists in presenting two hypotheses [5] (see Sections 4.1.2 and 4.2.1) and one novel approach as an alternative for variable selection in multivariate calibration (see Section 4.6).

3.5 Some Alternatives

As claimed by last section, genetic operators from standard GAs can cause schemata disruption. Currently, there are some approaches trying to deal with this issue. One alternative is known as competent genetic algorithms [50, 52, 98] (or Distribution Estimation Algorithms), which make use of modern strategies trying to avoid the disruption of BBs. Other approaches consist in the use of linkage learning and epistasis, which are genetic processes used to group and analyze dependent genes subset.

3.5.1 Compact genetic algorithm

Harik *et al.* [52] proposed the Compact Genetic Algorithm (CGA). CGA is an evolutionary strategy that mimics the order-one behavior of the standard GA. Whereas the standard GA adopts an initial set of solutions (population) and uses the recombination and mutation operators, CGA uses a probability vector $\mathbf{p} = [p_1 \quad p_2 \quad \ldots \quad p_l]^T$ in order to reproduce the offspring of the next generation.

The probability vector $\mathbf{p}$ has length l and its initial elements are equal to 0.5: $\mathbf{p} = [0.5 \quad 0.5 \quad \ldots \quad 0.5]^T$. Algorithm 3.1 shows a pseudocode for CGA.

At each generation, two individuals (**a** and **b**) are randomly generated from **p**, and a single tournament is performed between them in a fitness function. Then, the individual with the best performance is utilized to update **p**. As the algorithm evolves, **p** converges to an explicit solution

[5] Hypothesis 2 (Section 4.2.1) points out that the formation of BBs becomes unlikely in the presence of high data dimensionality.

Algorithm 3.1: Compact Genetic Algorithm, adapted from [52].

1: Let n be the population size, and l be the chromosome length
2: Initialize probability vector
3: Generate two individuals from the vector
4: Let them compete
5: Update the probability vector towards the better one
6: Check if the vector has converged
7: Vector $\mathbf{p}$ represents the final solution

$$\mathbf{p}(i) = \begin{cases} \mathbf{p}(i) + 1/l, & if \quad a(i) = 1 \quad and \quad a(i) \neq b(i), \\ \mathbf{p}(i) - 1/l, & if \quad a(i) = 0 \quad and \quad a(i) \neq b(i), \\ \mathbf{p}(i), & if \quad a(i) = b(i), \end{cases} \tag{3-6}$$

where $0 \leq i \leq l$.

Using such strategy, CGA requires less memory and performs faster and more efficient than standard GA when applied to easy problems [120]. However, when it is applied to difficult problems with high order schemata, CGA may not provide acceptable solutions and may be computationally infeasible for problems involving a relatively large number of variables [22]. It occurs because whereas GA evaluates the individual in the larger context, CGA do it bit to bit. Thus, CGA completely decorrelates the population's genes, which are usually interdependent [1, 36, 52].

Moreover, the update of probability vector proposed by Harik *et al.* [52] $(1/l)$ is quite simple because only the length (l) of the vector is considered. Empirically, it is possible to observe that a larger update step can make the probability vector converge quickly to mostly not acceptable results. On the other hand, a smaller update step increases the probability of $\mathbf{p}(i)$ to become zero, since the optimum is 1 (or vice versa) [28].

3.5.2 Linkage learning

Linkage learning (LL) is a genetic process grouping together dependent genes [6] subsets in GAs. It is based on genetic learning, which is a measure of distance between the loci of two or more specific dependent genes (or dependent/independent variables) [73]. According to Newman [73], many GA implementations that learn linkage is defined to operate by using a two phases method. Initially, they determine the subsets of dependent variables. Then, they use some strategy to improve the genetic linkage among such variables. Regarding LL, GAs have advantages over other search methods such as calculus-

[6] A gene is also referred as allele in some works from literature.

based methods, which provide them greater robustness and turn them capable of handling parameters made up of decision variable sets [51].

One of the ways chromosomes can populate the next generation is through crossover. It is known that when crossover is being performed, the groups of variables (BB) directly affecting the fitness of the chromosomes are usually split up. Considering crossover operator, the probability of splitting up a BB is proportionate to the defining length of such BB (see Section 3.4). On the other hand, if one of the parents was generated by an optimal (or near optimal) BB, the application of LL may become advantageous because it can reduce the probability of the BB being disrupted [17]. The generation of chromosomes with good linkage is advantageous because it helps to maintain viable subsets of variables [50]. In this case, independent variables could be placed adjacent to each other in order to turn the search more robust.

LL has been used to solve several problems (specially problems that standard GAs struggle to solve) due to the fact it can reduce the probability of BBs being broken up. However, LL in GAs (LLGA [51]) does not perform well on uniformly scaled problems [16]. The functional (or linear) dependency among variables is generally not known and hard to find, which means the aligning of independent variables adjacently must occur during the search for the optimal solution [73].

In addition, LLGA requires an exponentially-growing population size in terms of the number of BBs to solve an uniformly scaled problem [50]. Finally, there are some open questions which cast doubt on the use of LL. For example, why the LLGA behavior seems inconsistent when solving multiple BBs of different scalings? What conditions are required for a LLGA to learn linkage? And how independent variables can be aligned adjacently? These questions are not widely answered [73]. As far as we know, a few works have presented LLGA approaches to solve different types of problems [16, 51, 73], but none of them is specifically applied for variable selection in multivariate calibration.

3.5.3 Epistasis

Epistasis is a biological term that states the amount of intrachromosome gene interaction. In other words, it is a measure of interdependence between genes and an indicator of problem difficulty in GAs. According to Davidor [26], gene interaction is a central issue in natural genetics. Genes are usually dependent on each other and they can suppress as well as activate the expression of other genes.

When epistasis of a certain chromosome is high, many genes are strongly linked to other genes. It means that there is a strong correlation among genes in the chromosome [7]. In GA, epistasis is commonly used to indicate interdependency among

[7] In the context of this work, the high level of correlation among the variables in the dataset implies in a

the elements composing the representation. Representation is the primary aspect of a GA application and determines its performance [26].

Tracing epistasis is a hard task because the presence of epistatic elements (genes) can be traced only at the phenotypic level (representation) but not at the genotypic level (interaction) [15]. Even if the amount of epistasis is known, finding the epistatic genes and the level of interaction among them is considerably difficult. This issue can directly imply in the computational performance of the algorithm especially when dealing with high dimensionality problems. Furthermore, the more epistatic a given problem is the harder it may be for a GA to find its optimum solution [36].

Some researchers have proposed theoretical and empirical studies for epistasis measure. For instance, Davidor [26] is a pioneer in this area and presented an epistatic variance for estimating the epistasis degree in binary-coded representation GAs. Chan *et al.* [15] demonstrated how to use the approach of variance analysis to estimate the epistasis in real-coded representation GAs. Fonlupt *et al.* [36] proposed a bit-wise epistasis to measure the effect on the level of gene interaction for binary search spaces and improve the performance of evolutionary algorithms.

It is possible to check in such studies that strong epistatic relation of a representation is due to some of the bits. However, those proposed measures are able to estimate only the epistasis in the whole representation rather than estimating in each bit of the representation. Finally, as in the case of linkage learning and as far as we know, there is still no work in literature using epistasis measure in binary-coded representation GAs for selecting variables in multivariate calibration problems.

3.6 Summary

This chapter provided a detailed discussion about the building blocks (BBs) hypothesis. BB is a short and low-order schema with above average fitness. Schema is a template representing a set of solutions to a specific subproblem. Schema theory allows the definition of a BB and a statement of the BB hypothesis. Schema theorem aims to explain the Genetic Algorithm (GA) behavior. Basically, it describes how the number of copies of a specific schema depends on the selection and recombination processes when a standard GA is being utilized. Although schema theorem and its generalization have been cited by several works, both are restricted to some issues preventing them to provide an accurate analysis.

GAs have been widely used for variable selection in many optimization problems. However, some studies have emphasized that crossover operators used by

high level of epistatic relation (see Figures 6.1 and 6.5).

recombination-based search methods tend to cause the disruption of BBs. For example, in the context of multivariate calibration, we believe spectral regions representing the property of interest in the analyzed sample may be set in blocks and such blocks are often broken due to schemata disruption. Moreover, deceptive functions are considered as difficult problems to be solved by GAs, and if a problem is not (properly) decomposable, standard GA may not be able to find viable solutions.

The use of competent GAs and linkage learning can provide possible solutions for avoiding schemata disruption. However, there are still some issues (open questions) preventing linkage learning to be generalized and applied to some optimization problems solution. Despite the use of competent GAs to avoid the inconvenience of BBs disruption, such algorithms are computationally infeasible for problems involving large number of variables (high data dimensionality).

Finally, epistasis is an important concept used to measure the genes interdependence in the chromosome. Being aware about the gene interaction is a feasible manner to choose the best individuals of a GA to achieve the best outcomes. However, it is difficult to determine the epistatic genes and this may lead to performance decreasing in problems with high data dimensionality. In this context, Chapter 3 aimed to introduce all fundamental concepts in order to prepare the reader for the next one, which presents our proposal.

CHAPTER 4

Proposal

Indeed, multicollinearity is a concern (see Chapter 2). When two variables are correlated, one variable may carry information from the other one. Therefore, the presence of one without the other in the same subset of variables results in loss of relevant information [56]. In other words, when a dependent variable is selected, the absence of the variable to which it linearly depends may imply in the reduction of the model predictive ability [6, 39]. Consequently, multicollinearity also directly implies in the problem decomposition [1].

This chapter presents our two hypotheses. The first one (Section 4.1.2) deals with decomposability. It assumes the variable selection procedure in multivariate calibration can not be properly decomposed due to multicollinearity. One equation (based on works from literature) together with three numerical examples aim to claim its veracity.

The second hypothesis (Section 4.2.1) approaches the schemata disruption problem in Genetic Algorithms (GAs). It claims the schemata disruption caused by crossover operator directly affects the non-decomposability assumption from the first hypothesis. One proposition (also based on previous works) and three additional numerical examples aim to testify its viability.

This chapter also presents some algorithms implementations (Sections 4.3, 4.4, 4.5 and 4.6). One GA together with two heuristics and one local search operator are proposed. Additionally, we are presenting two versions of a feature selection algorithm based on epistasis concepts in order to avoid schemata disruption by preventing the problem decomposition.

4.1 Decomposability

Decomposability describes how the problem can be decomposed into smaller independent subproblems [53]. If there are subsets of variables which can be independently set of each other, the decomposability of a problem is considered high. Nevertheless, if

[1] If multicollinearity is high, the problem decomposition tends to be low [56, 89, 94].

a problem can not be decomposed into subproblems with few interdependencies among subsets of variables, the decomposability is low [94]. For instance, consider a random optimization problem with two variants. The first one assumes no decomposition of the problem. In this case, one decision variable can be defined as $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \mathbf{s}_4, \mathbf{s}_5, \mathbf{s}_6\}$. Consequently, there are six different solutions and the optimization method to be used must evaluate all possible solutions to find the optimal one.

The second variant assumes the problem can be decomposed. Then, two decision variables can be randomly chosen as $\mathbf{S}_1 = \{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3\}$ and $\mathbf{S}_2 = \{\mathbf{s}_4, \mathbf{s}_5, \mathbf{s}_6\}$. As shown in Table 4.1, a possible decomposition for such problem is $f = f_1(\mathbf{S}_1) + f_2(\mathbf{S}_2)$. It is possible to see that the resulting objective values f of different solutions are the same for both formulations. However, the second variant provides a lower size of the resultant search space. Therefore, the problem becomes easier for recombination-based search algorithms as the additive problem decomposition fits well the problem properties [94].

Table 4.1: *Two different formulations of a same problem, adapted from [94].*

Without problem decomposition $f = f(\mathbf{S}_1, \mathbf{S}_2)$	**Additive problem decompostion** $f = f_1(\mathbf{S}_1) + f_2(\mathbf{S}_2)$
$f(\mathbf{s}_1, \mathbf{s}_4) = 1, f(\mathbf{s}_1, \mathbf{s}_5) = 2, f(\mathbf{s}_1, \mathbf{s}_6) = 3$	$f_1(\mathbf{s}_1) = 0, f_2(\mathbf{s}_4) = 1$
$f(\mathbf{s}_2, \mathbf{s}_4) = 2, f(\mathbf{s}_2, \mathbf{s}_5) = 3, f(\mathbf{s}_2, \mathbf{s}_6) = 4$	$f_1(\mathbf{s}_2) = 1, f_2(\mathbf{s}_5) = 2$
$f(\mathbf{s}_3, \mathbf{s}_4) = 3, f(\mathbf{s}_3, \mathbf{s}_5) = 4, f(\mathbf{s}_3, \mathbf{s}_6) = 5$	$f_1(\mathbf{s}_3) = 2, f_2(\mathbf{s}_6) = 3$

4.1.1 Additive problem decomposition

According to Ahn [2], decomposable problems can be solved by concatenating basis functions of a certain order. Additively decomposable functions are one of the representations of a decomposable problem [2]. It follows that the overall fitness could be equal to the sum of all basis functions [68]. To demonstrate a problem is decomposable, Rothlauf [94] states that one can calculate the fitness of each variable separately and then calculate the $fitness_{Full}$ (overall fitness function) with all inputs together. If $fitness_{Full}$ is not equal to the sum of the variable fitnesses, it means such problem can not be properly decomposed [89].

In this sense, given the matrix $\mathbf{X}_{n \times k}$ from Equation (2-2) in Chapter 2, it could be decomposed if the objective value of each variable is individually calculated and its overall fitness function is obtained such as:

[2] When the problem is not decomposable, the sum of the fitness function of each subproblem will not equal the fitness of the whole problem.

$$f(\mathbf{X}_{n\times k}) = \sum_{i=1}^{k} f(\mathbf{x}^i_{n\times 1}). \tag{4-1}$$

where $\mathbf{X}_{n\times k} = \{\mathbf{x}^1_{n\times 1}, \mathbf{x}^2_{n\times 1}, ..., \mathbf{x}^k_{n\times 1}\}$, and the fitness function f could be assumed as the RMSEP (see Equation (2-3)).

Assuming the variable selection procedure in multivariate calibration as a decomposable problem and the use of a GA to select the variables, an additively decomposable function applied in this context could be formally defined as

$$f(\mathbf{X}_{n\times k}) = \sum_{i=1}^{N} f_{\mathbf{v}^i_{1\times k}}(\mathbf{X}^i_{n\times m}), \forall\ \mathbf{v}^i_{1\times k} \in \mathbf{V}_{N\times k}, \tag{4-2}$$

where $\mathbf{X}_{n\times k}$ is the matrix from Equation (2-2); $\mathbf{X}^i_{n\times m}$, $1 \leq m \leq k$, are N different subsets [3] of variables (columns) from $\mathbf{X}_{n\times k}$; $\mathbf{v}^i_{1\times k}$ is an individual from the population $\mathbf{V}_{N\times k}$ of the GA; and N is the number of individuals in $\mathbf{V}_{N\times k}$. More specifically, every $\mathbf{v}^i_{1\times k}$ is an $1 \times k$ 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.

Then, the variable subsets could be divided into several smaller subsets and recombined by genetic operators to form new better individuals. However, it is known a subset of informative variables may provide better outcomes than these variables divided into several subsets [6, 56]. Thus, such subset should not be split into different parts.

Additionally, when a problem can not be broken into smaller subproblems or its pieces affect one another (*i.e.*, interdependent subproblems), the problem can not be properly decomposed [2, 89, 94]. In this case, interdependent subproblems may contain information from other subproblems. Consequently, a partition between them may interfere in the final result [41, 94].

4.1.2 Hypothesis 1

The variable selection procedure in the context of multivariate calibration is not necessarily decomposable because such problem usually present high correlation degree among variables (see Section 2.2) [6, 84, 85, 101, 105]. As a consequence, the objective value (*e.g.*, RMSEP) of only one variable may be considerably worse than the RMSEP obtained from a subset of variables [39]. In this context, our first hypothesis arises:

- **Hypothesis 1**: Variable selection in multivariate calibration should be considered as a non-completely decomposable problem due to the considerable data correlation (multicollinearity) usually present in the dataset [83].

[3] It is important to note m is determined by the number of variables each individual selects. In other words, m is equal to the number of 1s of each separate individual in the GA population.

In order to claim the suitability of our first hypothesis, we are providing three numerical examples based on concepts of decomposability. The goal is to show the proposed Equation (4-2) can not be satisfied in the context of variable selection in multivariate calibration considering the standard GA with binary representation as variable selector. It is important to highlight that such hypothesis and examples are already published in the Proceedings of the Genetic and Evolutionary Computation Conference 2017 [83].

4.1.3 Numerical examples for hypothesis 1

Example 1. Let $\mathbf{X}_{n\times 4} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ be a subset with four random variables, where each $\mathbf{x}_i$, $1 \leq i \leq 4$, is an $n \times 1$ vector:

$$\mathbf{X}_{n\times 4} = \begin{bmatrix} -0.0023 & 0.0013 & -0.0022 & -0.0013 \\ -0.0025 & 0.0014 & -0.0023 & -0.0014 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ -0.0020 & 0.0010 & -0.0020 & -0.0012 \end{bmatrix}.$$

Moreover, let $\mathbf{V}_{4\times 4} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}$ be a population of a GA with four random individuals, where each $\mathbf{v}_i$, $1 \leq i \leq 4$, is an 1×4 vector:

$$\mathbf{V}_{4\times 4} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where $\mathbf{v}_1$ is the first row and selects the first column in $\mathbf{X}_{n\times 4}$, ..., and $\mathbf{v}_4$ is the last row selecting the last column in $\mathbf{X}_{n\times 4}$.

Calculating the prediction error (RMSEP) as fitness function for each separate variable, it is possible to obtain different individual values. Table 4.2 shows the RMSEP value for each variable in $\mathbf{X}_{n\times 4}$. Adding together all RMSEP values in Table 4.2, we obtain $f_{\mathbf{v}_1}(\mathbf{x}_1) + f_{\mathbf{v}_2}(\mathbf{x}_2) + f_{\mathbf{v}_3}(\mathbf{x}_3) + f_{\mathbf{v}_4}(\mathbf{x}_4) = 27.3339$.

Table 4.2: *RMSEP values for each variable in* $\mathbf{X}_{n\times 4}$.

Variable subset	RMSEP
$\mathbf{v}_1 \rightarrow \mathbf{x}_1$	10.7114
$\mathbf{v}_2 \rightarrow \mathbf{x}_2$	3.9159
$\mathbf{v}_3 \rightarrow \mathbf{x}_3$	6.1204
$\mathbf{v}_4 \rightarrow \mathbf{x}_4$	6.5862

Example 2. Consider now $\mathbf{V}_{8\times 4} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4, \mathbf{v}_5, \mathbf{v}_6, \mathbf{v}_7, \mathbf{v}_8\}$ as a population with eight random individuals:

$$\mathbf{V}_{8\times4} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

where each individual selects different combinations of variables. For instance, $\mathbf{v}_1$ selects $\{\mathbf{x}_1, \mathbf{x}_2\}$, ..., and $\mathbf{v}_8$ selects $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$.

Again, calculating the fitness function for each individual it is possible to obtain different values. Table 4.3 shows the separately-obtained RMSEP values by different combinations of variables in $\mathbf{X}_{n\times4}$. Note the RMSEP value for the four variables together in the same subset is $f_{\mathbf{v}_8}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = 3.4937$.

Table 4.3: *RMSEP values for different combinations of variables in matrix* $\mathbf{X}_{n\times4}$.

Variable subset	RMSEP
$\mathbf{v}_1 \rightarrow \{\mathbf{x}_1, \mathbf{x}_2\}$	3.9415
$\mathbf{v}_2 \rightarrow \{\mathbf{x}_1, \mathbf{x}_3\}$	6.1242
$\mathbf{v}_3 \rightarrow \{\mathbf{x}_1, \mathbf{x}_4\}$	6.5756
$\mathbf{v}_4 \rightarrow \{\mathbf{x}_2, \mathbf{x}_3\}$	3.8765
$\mathbf{v}_5 \rightarrow \{\mathbf{x}_2, \mathbf{x}_4\}$	3.9110
$\mathbf{v}_6 \rightarrow \{\mathbf{x}_3, \mathbf{x}_4\}$	5.8188
$\mathbf{v}_7 \rightarrow \{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$	3.4252
$\mathbf{v}_8 \rightarrow \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$	3.4937

Example 3. Calculating the Pearson's linear correlation coefficient (see Section 2.4.3) for matrix $\mathbf{X}_{n\times4}$, one can obtain a symmetric matrix $\mathbf{R}_{4\times4}$ [4] such as:

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

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 $\rho_{\mathbf{x}_2,\mathbf{x}_3}$ = -0.0099 (close to zero), which means they are near linearly independent. Nevertheless,

[4] Matrix $\mathbf{R}$ is related to the covariance of matrix $\mathbf{X}$. In matrix $\mathbf{R}$, all main diagonal elements are equal to 1, which means every variable is directly correlated to itself.

variables $\mathbf{x}_3$ and $\mathbf{x}_4$ have $\rho_{\mathbf{x}_3,\mathbf{x}_4} = 0.9843$ (close to 1) and both are (almost totally) correlated. Hence, $\mathbf{x}_3$ and $\mathbf{x}_4$ are correlated to each other such that one variable may carry information from the other (*e.g.*, epistatic relation - see Section 3.5.3).

We can notice the RMSEP values sum of each separate variable in Table 4.2 ($f_{\mathbf{v}_1}(\mathbf{x}_1) + f_{\mathbf{v}_2}(\mathbf{x}_2) + f_{\mathbf{v}_3}(\mathbf{x}_3) + f_{\mathbf{v}_4}(\mathbf{x}_4) = 27.3339$) is considerably greater than the obtained RMSEP value with the four variables together in Table 4.3 ($f_{\mathbf{v}_8}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = 3.4937$). Then, it is possible to claim $f_{\mathbf{v}_8}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) \neq f_{\mathbf{v}_1}(\mathbf{x}_1) + f_{\mathbf{v}_2}(\mathbf{x}_2) + f_{\mathbf{v}_3}(\mathbf{x}_3) + f_{\mathbf{v}_4}(\mathbf{x}_4)$, which means the presence of interdependent subset of variables indeed does not allow a proper decomposition of the problem [93].

Additionally, the subset composed of variables $\{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ in Table 4.3 provides the lowest RMSEP ($f_{\mathbf{v}_7}(\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = 3.4252$). 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 $\mathbf{x}_3$ ($f_{\mathbf{v}_4}(\mathbf{x}_2, \mathbf{x}_3) = 3.8765$) and $\mathbf{x}_4$ ($f_{\mathbf{v}_5}(\mathbf{x}_2, \mathbf{x}_4) = 3.9110$), these three variables are able to reduce the prediction error since they remain together in the same subset.

Those three numerical examples indicate Equation (4-2) can not be satisfied for the variable selection procedure in multivariate calibration. Therefore, they provide significant evidences that selecting variables in multivariate calibration should indeed be considered as a non-completely decomposable problem, which supports our first hypothesis [83].

Finally, Section 4.1.4 briefly discuss the polynomial decomposition. It is an approach developed in the literature to estimate how well a problem can be solved by using recombination-based search methods. Such approach assumes search performance is higher if the problem can be decomposed into smaller subproblems [93].

4.1.4 Polynomial problem decomposition

Rothlauf [94] states that the linearity of an optimization problem can be measured by its polynomial decomposition. For binary decision variables, any objective function f defined on l decision variables $x_i \in \{0, 1\}$ can be decomposed according to Equation (4-3):

$$f(\mathbf{x}) = \sum_{i \subset \{1,\ldots,l\}} \alpha_i \prod_{j \in i} \mathbf{e}_j^T \mathbf{x}, \tag{4-3}$$

where the vector $\mathbf{e}_j$ contains 1 in the jth column and 0 elsewhere, T denotes transpose, and the α_i are the coefficients describing the non-linearity of the problem.

Regarding $\mathbf{x} = (x_1, \ldots, x_l)$, the function f may be viewed as a polynomial in the variables $x_1, \ldots, x_l$. If the decomposed problem has only order 1 coefficients, the problem is linear decomposable. However, if there are high order coefficients, then the

problem function is non-linear. The highest polynomial coefficient is able to determine the maximum non-linearity of the function f. Therefore, the higher the order of the α_i, the more non-linear the problem is [94].

According to Mason [67], there is some correlation between the non-linearity of a problem and its difficulty for recombination-based search methods. Nevertheless, there could be high order α_i although the problem can be easily solved by recombination-based search algorithms. For example, the function

$$f(\mathbf{x}) = \begin{cases} 1 & \text{for} \quad x_1 = x_2 = 0 \\ 2 & \text{for} \quad x_1 = 0, x_2 = 1 \\ 4 & \text{for} \quad x_1 = 1, x_2 = 0 \\ 10 & \text{for} \quad x_1 = x_2 = 1, \end{cases} \tag{4-4}$$

can be decomposed into $f(\mathbf{x}) = \alpha_1 + \alpha_2 x_1 + \alpha_3(x_2^2) + \alpha_4 x_1(x_2^2) = 1 + 3x_1 + (x_2^2) + 5x_1(x_2^2)$.

It is possible to see the problem is decomposable. Thus, it is easy for a GA as each of the two decision variables can be solved independently of each other. On the other hand, as the problem is non-linear and high order coefficients exist, the polynomial decomposition wrongly classifies the problem as difficult. This misclassification is due to the fact the polynomial decomposition assumes a linear decomposition and can not appropriately describe non-linear dependencies [67, 94]. In this case, the use of different approaches such as Walsh decomposition may be more viable [40]. Note that it is beyond the scope of this work to investigate and compare differences between problem decomposition techniques.

Considering decision variables in the context of multivariate calibration and Equation (4-3), an objective function f defined on k variables $\mathbf{x}_{n\times 1} \in \mathbf{X}_{n\times k}$ could be decomposed into [5]:

$$f(\mathbf{X}_{n\times k}) = \sum_{i=1}^{k} \beta_i \prod_{j=i} \mathbf{e}_j^T \mathbf{X}_{n\times k}, \tag{4-5}$$

where the vector $\mathbf{e}_j$ contains 1 in the jth column and 0 elsewhere. The β_i are the coefficients describing the (possible) non-linearity of the problem.

For instance, let $\mathbf{X}_{n\times 2} = \{\mathbf{x}_1, \mathbf{x}_2\}$ be a subset with two random columns from matrix $\mathbf{X}_{cal}$ (see Section 2.1), where each $\mathbf{x}_i$, $1 \leq i \leq 2$, is an $n \times 1$ vector:

[5] Equation (4-5) was adapted from Rothlauf [94].

$$\mathbf{X}_{n\times 2} = \begin{bmatrix} -0.0022 & -0.0013 \\ -0.0023 & -0.0014 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ -0.0020 & -0.0012 \end{bmatrix}.$$

Calculating the regression coefficients through Equation (2-2), it is possible to obtain

$$\beta = \begin{bmatrix} -2.0860\times 10^4 \\ 2.3254\times 10^4 \end{bmatrix}.$$

In this case (and usually in multivariate calibration), we do not have binary decision variables. Instead, $\mathbf{x}_1$ and $\mathbf{x}_2$ are two different linear-continuously decision variables and there are only order 1 coefficients with one possible outcome: $f(\mathbf{X}_{n\times 2}) = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 = 1 + (-2.0860\times 10^4)\mathbf{x}_1 + (2.3254\times 10^4)\mathbf{x}_2$. Therefore, this is an evidence that the problem might be linear decomposable if both variables are linearly independent (uncorrelated). However, variables $\mathbf{x}_1$ and $\mathbf{x}_2$ provide a Pearson's linear correlation $\rho_{\mathbf{x}_1,\mathbf{x}_2} = 0.9843$ [6], which indicates a high correlation degree between them and indeed the problem can not be properly decomposed.

As far as we know, related works in literature have not considered such decomposition issues accordingly. Many works have used different versions of GAs to select variables in multivariate calibration problems [6, 12, 37, 64, 79, 105, 124]. However, although obtaining viable outcomes, those algorithms have an inherently stochastic characteristic. Consequently, such characteristic is usually due to the use of genetic operators [7]. In this sense, Section 4.2.1 presents our second hypothesis. It approaches the schemata disruption problem caused by recombination operators in the context of multivariate calibration.

4.2 Schemata Disruption

Recombination-based search methods such as GAs aim to solve problems by trying different decompositions of the problem [94]. They try solving the resultant subproblems and putting together the obtained solutions to get a final solution to the overall problem. Solving a certain number of smaller subproblems is usually easier than

[6] Such value was calculated through Equation (2-12). See Example 3 in Section 4.1.3.

[7] One possible explanation is that recombination-based search methods exploit the search space by getting better and better individuals through crossover and mutation. While crossover aims to generate an offspring superior to the parents, mutation tries to escape from local optimum regions.

solving the original problem, and this leads to high performance of recombination-based search algorithms [93]. Thus, it is important to have no interactions among variables from different subsets so that the algorithm becomes able to select the most informative variables.

Although recombination-based search methods are operationally simple, they may be considered as complex algorithms [68]. To understand and design them on an efficient manner, one can treat them as a divide-and-conquer approach by decomposing the problem into smaller decomposable subproblems, solving those subproblems independently and integrating them as a whole solution. Therefore, in order to use a recombination-based optimization method accordingly, one should define the decision variables of a problem such that they allow a decomposition of the problem [93].

As mentioned before, to a problem to be decomposable there must be none interaction between any two variables and each variable should be separately treated [115]. A problem can be properly decomposed by identifying the interdependencies among different genes [8] in the chromosome of an individual from the population of a GA. In this sense, the purpose of genetic operators should consist in decomposing the problem by detecting which bits in the string influence each other (*e.g.*, epistasis). Nonetheless, detecting such interdependencies among the bits in the string may become a considerable complicating factor, especially when dealing with long string length and high order schemata [41].

Recombination operators are techniques used in GAs to decompose an optimization problem into smaller subproblems in order to generate new solutions at each generation. Each individual in a population can be considered as a possible solution. One of the most known recombination operators is the one-point crossover. One-point crossover is a simple operator which splits the subjects into two parts and aims to regroup different parts to form new individuals (offspring). The goal usually consists in obtaining an offspring better than parents. For example, if the objective value (fitness function) of the offspring is better than the parents fitness, then the offspring should be preserved in the population. However, splitting an individual commonly implies in an inconvenience called building blocks (or schemata) disruption [47].

A building block (BB) is a schema with low order, short defining-length and an above-average fitness (see Section 3.1). It is able to generate the best individuals and its characteristics must be inherited by new offsprings over generations. Notwithstanding, the use of recombination operators to split parents chromosome and form an offspring can directly cause the disruption of BBs [41, 47]. Thus, a disrupted BB is not preserved in new individuals and the GA's performance tends to decrease in next iterations [62].

[8] Genes may be considered as bits in a binary-coded string, and each bit may be treated as a variable position.

On the one hand, if the problem can be decomposed into smaller subproblems and the intra-BB and inter-BB difficulty (see Section 3.4.1) are low, such problem may be considered as an easy one for recombination-based search methods [41]. The decomposition is based on the assumption that such methods can decompose problems and work with BBs accordingly [93]. Goldberg *et al.* [43] claim the decomposition of a problem consists in the following conditions:

- It is important to know the algorithm process, which means considering the BBs processing;
- Solving tractable problems by BBs;
- Supplying enough BBs in the initial population;
- Ensuring the growth of viable BBs;
- Mixing BBs properly;
- Deciding between two competing BBs.

On the other hand, considering all these six points, it becomes clear how great is the complexity of a recombination-based search algorithm to ensure a good solution (*e.g.*, global optimum) through a viable search space exploration.

GAs have been successfully used to select variables in several optimization problems [5, 64, 85, 101, 121]. Even when decomposability of the problem is low, GAs make use of recombination operators to split it into subproblems and integrate them as a whole solution. However, examples from Section 4.1.2 show us that breaking the problem into independent subsets of variables is different from dividing these subsets into interdependent smaller parts [83, 84].

4.2.1 Hypothesis 2

When dealing with a problem which can not be properly decomposed, the use of recombination-based search methods may not be helpful as no suitable decomposition of the problem is possible [94]. In this case, efforts of such methods to find proper decompositions may become useless [93]. Therefore, considering the schemata disruption problem our second hypothesis arises:

- **Hypothesis 2**: Schemata disruption caused by recombination operators in GAs directly affects the non-decomposability assumption of the variable selection procedure in multivariate calibration [84].

To demonstrate the feasibility of our second hypothesis, we are providing three numerical examples based on concepts of schema theory. The goal is to show that not necessarily there exists building blocks formation in spectral data from multivariate

calibration due to the high data dimensionality. Even so, the crossover operator tends to cause schemata disruption, and this issue can affect the previous hypothesis (Section 4.1.2). It is important to highlight that an earlier version of the second hypothesis is already published in the Proceedings of the Genetic and Evolutionary Computation Conference 2016 [84].

4.2.2 Numerical examples for hypothesis 2

As thoroughly discussed in Section 3.4, BBs are low-order and short defining-length schemata with an above-average fitness. Thus, consider the following proposition based on Chung's statement [19]:

Proposition 1 (For a schema to be considered as a building block). *If these three conditions are not satisfied, then a given schema H can not be considered as a BB:*

1. *The number of fixed positions, $O(H)$, is low (low-order);*
2. *The distance between the two outermost fixed bits, $\delta(H)$, is short (short defining-length);*
3. *The fitness of schema H, $f(H)$, is better than the average fitness of the population (above-average fitness).*

Example 1. Consider H_1 = 101*011 and H_2 = 1*****0 as two random schemata. It is easy to see schema H_1 has not a low-order ($O(H_1) = 6$) because most of positions are fixed bits, and has not a short defining-length ($\delta(H_1) = 6$) because the distance between the two outermost fixed bits is as large as possible.

Schema H_2 has a low-order ($O(H_2) = 2$) but has not a short defining-length ($\delta(H_2) = 6$). Consequently, according to Proposition 1 both schemata could not be considered as BBs. It is important to note the verification of condition 3 of Proposition 1 is problem-dependent. In other words, it depends on the fitness function of the problem. Next example approaches all three conditions of Proposition 1.

Example 2. Considering Equation (4-2), let $\mathbf{X}_{n\times 4} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ be the same matrix from Example 1 in Section 4.1.3, where $\mathbf{x}_1$ is the first column, ..., and $\mathbf{x}_4$ the last column:

$$\mathbf{X}_{n\times 4} = \begin{bmatrix} -0.0023 & 0.0013 & -0.0022 & -0.0013 \\ -0.0025 & 0.0014 & -0.0023 & -0.0014 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ -0.0020 & 0.0010 & -0.0020 & -0.0012 \end{bmatrix}.$$

Let $\mathbf{V}_{4\times4} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}$ be the population of a GA with four different individuals (subjects):

$$\mathbf{V}_{4\times4} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix},$$

where $\mathbf{v}_1$ is the first row, ..., and $\mathbf{v}_4$ is the last row. Each subject selects different combinations of variables (columns) in $\mathbf{X}_{n\times4}$. For instance, $\mathbf{v}_1$ selects $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$, ..., and $\mathbf{v}_4$ selects $\{\mathbf{x}_2, \mathbf{x}_4\}$.

Moreover, let $H_1 = 1***$ be a schema for subjects $\mathbf{v}_1$ and $\mathbf{v}_3$ in $\mathbf{V}_{4\times4}$. For H_1 to be considered as a BB, it is known that all three conditions of Proposition 1 must be satisfied. Schema H_1 has low-order ($O(H_1) = 1$) and a short defining-length ($\delta(H_1) = 0$). Using Equation (3-1) in Section 3.1 to calculate the fitness of schema H_1, we have [9]:

$$f(H_1) = \frac{f_{\mathbf{v}_1}(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4) + f_{\mathbf{v}_3}(\mathbf{x}_1)}{2} = \frac{3.4937+10.7114}{2} = 7.1025.$$

Calculating the average fitness of all subjects in $\mathbf{V}_{4\times4}$, we have:

$$f(\mathbf{V}_{4\times4}) = \frac{f_{\mathbf{v}_1}(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4) + f_{\mathbf{v}_2}(\mathbf{x}_2) + f_{\mathbf{v}_3}(\mathbf{x}_1) + f_{\mathbf{v}_4}(\mathbf{x}_2,\mathbf{x}_4)}{4} = \frac{3.4937+3.9159+10.7114+3.9110}{4} = 5.5080.$$

Despite schema H_1 has low-order and a short defining-length, it is possible to see $f(H_1) > f(\mathbf{V}_{4\times4})$. Since selecting variables in multivariate calibration often aims to reduce the fitness function value (*i.e.*, minimization problem), it is clear H_1 has not an above-average fitness and indeed can not be considered as a BB because condition 3 of Proposition 1 is not satisfied.

Example 3. Assuming the existence of BBs in spectral data from multivariate calibration, consider $\mathbf{v}_1$ (subject 1) and $\mathbf{v}_2$ (subject 2) from $\mathbf{V}_{4\times4}$ (last example) in Figure 4.1 generated by schema $H_1 = **11$ and schema $H_2 = 01**$, respectively. In the reproduction process between subject 1 and subject 2, two new subjects are obtained.

In this example, schema H_1 can generate only $\mathbf{v}_1$ in $\mathbf{V}_{4\times4}$. It has $O(H_1) = 2$; and $\delta(H_1) = 1$. Schema H_2 can generate subject $\mathbf{v}_2$ as well as subject $\mathbf{v}_4$. It has $O(H_2) = 2$; and $\delta(H_2) = 1$. Calculating their fitness function through Equation (3-1), we can obtain [10]:

$$f(H_1) = \frac{f_{\mathbf{v}_1}(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4)}{1} = \frac{3.4937}{1} = 3.4937.$$

$$f(H_2) = \frac{f_{\mathbf{v}_2}(\mathbf{x}_2) + f_{\mathbf{v}_4}(\mathbf{x}_2,\mathbf{x}_4)}{2} = \frac{3.9159+3.9110}{2} = 3.9134.$$

[9] Note the values were obtained from Tables 4.2 and 4.3 in Section 4.1.3.
[10] Values also obtained from Tables 4.2 and 4.3.

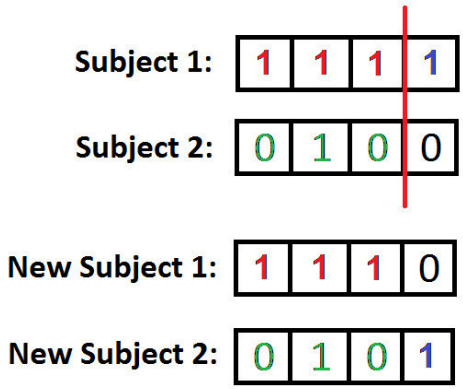

Figure 4.1: *Example of an offspring generated from two individuals by one-point crossover.*

Both schemata have short defining-length and are above-average fitnesses ($f(H_1) < f(\mathbf{V}_{4\times4})$ and $f(H_2) < f(\mathbf{V}_{4\times4})$). Considering H_1 and H_2 as low-order [11], then the three conditions of Proposition 1 are satisfied and they both could be considered as BBs. However, this example approaches a simple case with an extremely reduced subset of variables. On the other hand, in this work we deal with two large datasets (see Chapter 5 for more details). Thereby, the existence of BBs in this type of data can not be assured due to the presence of high data dimensionality [41, 93, 94]. Even not selecting all variables, as any individual in the GA population can have large string length, then a given schema H:

- may not have a small number of fixed positions, which would not satisfy condition 1 of Proposition 1;
- may not have a short defining-length, which would not satisfy condition 2 of Proposition 1; and
- may not have an above-average fitness, which would not satisfy condition 3 of Proposition 1.

In other words, considering Proposition 1 it is not possible to guarantee the existence of BBs in datasets with high data dimensionality [12]. In this case, the schemata have large string lengths. Thus, such schemata tend to have a large defining-length, which does not satisfy condition 2 of Proposition 1. Moreover, the number of fixed positions in the schemata tends to be large due to long distances between the first and last fixed bits, which may increase the amount of fixed bits (not satisfying condition 1 of Proposition 1). Finally, as stated earlier condition 3 of Proposition 1 depends on the fitness function.

[11] Half of their positions are fixed, but it is not the most of them.

[12] According to Proposition 1, every building block is a schema but not all schema is considered as a building block.

In addition, as crossover dropped down exactly between the two fixed bits in subject 1 from Figure 4.1, one schema disruption occurs. It is possible to notice it in the new subject 1, which means schema H_1 is not preserved in the next generation. Then, significant performance may be lost if H_1 should be preserved. For example, Table 4.3 from Section 4.1.3 shows us variables $\mathbf{x}_3$ and $\mathbf{x}_4$ provide the lowest RMSEP value when they are together in the same subset with variable $\mathbf{x}_2$. Therefore, this indicates H_1 should indeed be inherited by next generations but it does not happen.

According to Goldberg [41], the probability of schemata disruption under crossover is given by Equation (4-6). This is due to the fact a schema is disrupted every time crossover operator falls within its defining-length [47].

$$P_{disruption}(H) = \frac{\delta(H)}{l-1}, \tag{4-6}$$

where H is a given schema, $\delta(H)$ is the defining-length of schema H, and l is its string length.

Although schema H_1 and schema H_2 in this example may be considered as BBs, their disruption probabilities are the same:

$$P_{disruption}(H_1) = \frac{\delta(H_1)}{l-1} = \frac{1}{3} = 0.3333.$$

$$P_{disruption}(H_2) = \frac{\delta(H_2)}{l-1} = \frac{1}{3} = 0.3333.$$

Schema H_1 and schema H_2 have approximately 33% of probability to be disrupted since crossover operator has three possible places to cross the chromosome. Then, the disruption probability of a given schema is directly proportional to its defining-length [41]. As a consequence, the greater the size of defining-length, the greater the probability of disruption. Therefore, such example also indicates schemata may be commonly disrupted in our large datasets (see Sections 5.1 and 5.2) [84].

Finally, another issue may arise. As described earlier, one single variable may not contribute with some relevant information but it may become important when it is in conjunction with other variables [18, 56]. Similarly, if two variables are correlated, they should remain together in the same subset. The presence of one of them without the other may negatively affect the fitness value. Thus, assuming the third ($\mathbf{x}_3$) and fourth ($\mathbf{x}_4$) variables in subject 2 from Figure 4.1 contribute to obtain a more reduced RMSEP value when they are together or absent in a same subset, subjects from next generation will not inherit such important characteristic either. This inconvenience also affects the non-decomposability assumption approached by Hypothesis 1 [83]. Consequently, significant

information may be lost through crossing operations between two individuals of the population even when a schema disruption does not occur [13].

Such three examples indicate Proposition 1 may not be totally satisfied for the variable selection procedure in multivariate calibration. Therefore, the BBs formation becomes unlikely in high dimensionality problems. Furthermore, all examples provide considerable evidences about schemata disruption in this context as well as it affects the non-decomposability assumption of the problem, which supports our second hypothesis [84].

As far as we know, related works in literature have not considered this issue either [37, 64, 79, 87, 102, 103, 104]. Although being possible to use competent GAs to avoid schemata disruption (see Section 3.5), they are not suitable for all types of representation and may demand a significant computational effort when dealing with large datasets [41, 50, 99]. In this sense, a hybridization between concepts of compact GA (Section 3.5.1) and epistasis (Section 3.5.3) may be a better alternative to tackle all those inconveniences (see Section 4.6). Next sections present our proposed algorithms.

4.3 Genetic Algorithm Implementation

Following implementations from some main works in literature which have used GAs for variable selection in multivariate calibration [5, 61, 74], our GA uses a chromosome with a fixed number of genes (as many as the variables), each one of them being just 1 bit long (0 = absent variable; 1 = present variable). Crossover operator has been used [14]. Algorithm 4.1 shows the pseudocode for the GA implementation.

In Algorithm 4.1, fitness function consists in the number of selected variables by each individual. Thus, solutions are evaluated based on the number of selected variables. Parameter p_c ($0 \leq p_c \leq 1$) is the probability of crossover occurrence. The terms *offspring_best* and *offspring_worst* consist in two counters which indicate, respectively, the number of children better and worse than parents. Moreover, all individuals are arranged in pairs (parents) and used to generate two new individuals. Finally, we use *rand* function to initialize the population of individuals. It is a *Matlab*© built-in function which yields an uniform random number.

[13] Even not happening a schema disruption, the problem decomposition may separate the variables into two or more interdependent subproblems.

[14] Different operators can be looked into based on the problem settings. However, in this study, we are focusing only on the use of one-point crossover as recombination operator.

Algorithm 4.1: Proposed GA implementation.

```
1: Let N be the population size
2: Initialize the population of individuals using uniformly distributed random numbers
   or some heuristic
3: for i = 1 : MaxGenerations
4:    Evaluate all individuals based on their fitness
5:    if rand < p_c
6:       Generate a fixed number of solutions using crossover operator
7:       if offspring fitness is better than parents fitness
8:          Replace parents by children increasing the number of offspring_best
9:       else
10:         Ignore children and increase the number of offspring_worst
11:      end if
12:   end if
13: end for i
```

4.4 Heuristics

In this work, we are proposing two heuristic strategies into the GA implementation: *i*) initial individuals generation for the proposed GA, which aims to select the best chromosomes (schemata candidates) as initial solutions; and *ii*) schemata identification, which aims to determine the possible schema generating the best chromosomes. In this case, best chromosomes are featured by the set of solutions providing the best fitness values (*e.g.*, smallest RMSEP). Finally, the best schema is determined by the fixed positions (genes set as 1) of the best chromosome over generations.

4.4.1 Heuristic for initial solutions generation

This heuristic applies a relatively simple strategy based on the Successive Projections Algorithm [6]. It aims to reduce the model prediction error (RMSEP) by initially selecting only (near) orthogonal variables (Sections 6.2.1 and 6.2.2 show some outcomes). Algorithm 4.2 shows the pseudocode for such heuristic [15].

In Algorithm 4.2, parameter m is the number of variables to be initially selected and N is the number of individuals in the GA population. First of all, it is obtained in step 3 a copy of matrix $\mathbf{X}$ into $\mathbf{X}_{projected}$, where $\mathbf{X}$ is an $n \times k$ matrix from the calibration set (see Chapter 5). Then, it is performed in step 4 the square norm of each column of matrix $\mathbf{X}_{projected}$. In other words, square norm is performed by adding the square of all elements from each column of $\mathbf{X}_{projected}$ and storing the sum outcomes in the row vector **norms**, where each element of **norms** represents the respective column of matrix $\mathbf{X}_{projected}$.

[15] It is important to note Algorithm 4.2 performs the step 2 of Algorithm 4.1.

Algorithm 4.2: Proposed heuristic to generate initial solutions.

1: **Input parameters:** $\mathbf{X}$ and m
2: **for** $i = 1 : N$
3: $\quad \mathbf{X}_{projected} = \mathbf{X}$
4: $\quad \mathbf{norms} = \text{sum}(\mathbf{X}^2_{projected})$
5: $\quad norm\_max = \max(\mathbf{norms})$
6: $\quad \mathbf{X}_{projected} = \frac{\mathbf{X}_{projected} * norm\_max}{\mathbf{norms}_i}$
7: $\quad [\mathbf{Q}, \mathbf{R}, \mathbf{order}] = \text{qr}(\mathbf{X}_{projected})$
8: $\quad \mathbf{chain} = \mathbf{order}(1 : m)$
9: $\quad \mathbf{Population}_{i,\mathbf{chain}} = 1$
10: **end for** i

Shortly thereafter, step 5 obtains the largest element from vector **norms**. Then, step 6 scales the i-th column of $\mathbf{X}_{projected}$ so that it becomes the largest column. Scaling the i-th column consists in multiplying it by the value obtained in the previous step and dividing it by the norm of column i.

Step 7 uses qr function to produce an orthogonal-triangular decomposition divided into two different matrix (**Q** and **R**). The qr is a $Matlab^{©}$ built-in function which produces an orthogonal-triangular decomposition by using mathematical operations. It can be used to provide an orthogonal basis of the column space of a matrix. In its general form, it produces an $n \times k$ upper triangular matrix **R** and an $n \times n$ orthogonal matrix **Q** so that $\mathbf{X}_{projected} = \mathbf{QR}$.

Vector **order** contains the order of the most orthogonal columns (variables) according to function qr. In this case, step 8 obtains the m first variables from such vector and stores them in vector **chain**. Finally, step 9 sets "1" in each gene of the i-th individual according to the index of each variable.

It is important to note the qr function uses reflections (projections operation) to compute matrix **Q** and obtain an orthogonal basis. Function qr does not create new variables. Instead, we are only using the pivoting of the qr function. In other words, it aims to select the most orthogonal variables based on mathematical projections operation [6, 105].

Finally, it is also important to emphasize this heuristic is not aware about the possible schemata formation or disruption. It only tries to select the best variables (most orthogonal to each other) from the dataset in order to initially compose the calibration model.

4.4.2 Heuristic for possible schemata identification

By using this second heuristic, we aim to identify the schema which possibly generates the best individuals in the GA population. Algorithm 4.3 provides the pseu-

docode for the possible schemata identification heuristic.

Algorithm 4.3: Proposed heuristic to identify possible schemata.

1: **for** $i = 1 : MaxGenerations$
2: Evaluate all individuals based on their fitness (RMSEP value)
3: Sort all individuals in ascending order by their fitness
4: Obtain the indexes of the best individuals
5: Assume that one schema is formed by the chromosome of the best individual in which "1" means a fixed position in the schema and the respective variable is to be selected
6: sum = fitness(best individual)
7: $count = 1$
8: **for** $j = 2$: number of best individuals
9: **if** schema is included in individual j
10: $sum = sum$ + fitness(individual j)
11: $count = count + 1$
12: **enf if**
13: **end for** j
14: schema_fitness = $\frac{sum}{count}$
15: **end for** i

In Algorithm 4.3, the fitness function consists in the RMSEP (Equation (2-3)) value obtained by each individual. Firstly, step 2 performs an individual evaluation and step 3 sort them from the best to the worse one. Then, step 4 obtains the indexes of the $x\%$ best individuals in the population, where x can be empirically chosen. Step 5 generates one matrix with those best individuals [16] and chooses the best individual as the one who provides the best fitness (smallest RMSEP) value assuming it as the schema which generates the better ones.

Step 6 obtains the fitness of the best individual from step 4. From step 7 to 11, it is checked which individuals may be generated by the schema. For this, we use the $Matlab^{©}$ buit-in function min(schema, individual j). Such function returns an array with the smallest elements taken from parameters schema or individual j. In other words, it verifies if individual j, $2 \leq j \leq$ number of best individuals, contains the same fixed bits (genes) from the schema. If so, then the individual fitness is added and a counter is incremented. Finally, step 14 performs Equation (3-1) and obtains the schema fitness.

It is relevant to note we are assuming genes "1" from the best individual [17] at each generation are fixed positions of a schema which possibly generates the best individuals (chromosomes with the most informative variables). Then, the fitness of all

[16] In such matrix, each row represents an individual and each column indicates the respective variable to be (or not) selected.

[17] In the context of this work, the best individual is the one who provides the smallest RMSEP value.

best individuals containing such schema is added so that the schema fitness can be calculated according to Equation (3-1).

By using such heuristic, we aim to analyze the possible schemata disruption over generations (see Section 6.2). Finally, it is also relevant to emphasize the strategy used in this heuristic is naive and it is not able to ensure the formation of schemata neither avoid their disruption.

4.5 Local Search-based Operator

As claimed in Section 4.2.1, recombination operators tend to cause the disruption of schemata in standard GA implementations. Considering this issue as an inconvenience when selecting variables in large datasets from multivariate calibration models, we are also proposing an operator which performs a local search in order to find better individuals (solutions) and reduce schemata disruption. This operator execute a simple local search by modifying particular genes (variables) in the chromosome. Algorithm 4.4 shows the pseudocode for the local search-based operator [18].

Algorithm 4.4: Proposed local search-based operator.

```
for i = 1 : number of individuals / 2
    Obtain individual 1 as father
    Obtain individual 2 as mother
    son = father
    daughter = mother
    j = 1
    Mutate gene j in son
    while fitness son is worse than fitness father
      Undo the previous mutation
      j = j + 1
      Mutate gene j in son
    end while
    j = 1
    Mutate gene j in daughter
    while fitness daughter is worse than fitness mother
      Undo the previous mutation
      j = j + 1
      Mutate gene j in daughter
    end while
    Replace father by son and mother by daughter
end for i
```

[18] It is important to note Algorithm 4.4 performs the step 6 of Algorithm 4.1 by using a simple local search instead of crossover operator.

In Algorithm 4.4, steps 2 and 3 select two different individuals from the population to participate in the reproduction process. Instead of using a recombination operator, steps 6-12 and 13-19 go through each gene of both chromosomes mutating it in order to obtain an offspring better than parents. Such algorithm performs a naive strategy and is based on the Variable-Neighborhood Search method [94].

It is noteworthy Algorithms 4.2, 4.3 and 4.4 serve as complement for Algorithm 4.1. Each one utilizes a different heuristic strategy aiming to explore the research empiricism. Their goal consists in improving the standard GA outcomes.

Finally, next section presents two versions of a novel approach for variable selection in multivariate calibration (the Epistasis-based Feature Selection Algorithm). Such algorithm uses two different enhanced strategies in order to avoid the schemata disruption by preventing the problem decomposition.

4.6 Epistasis-based Feature Selection Algorithm

Considering the issues raised up by Sections 4.1.2 (non-decomposability assumption) and 4.2.1 (schemata disruption hypothesis), this work is presenting an Epistasis-based Feature Selection Algorithm (EbFSA). According to Davidor [26], the effect of epistasis lies in the ability to predict the value of a whole from the value of its parts. However, finding epistatic genes is computationally difficult in high dimensionality problems [36]. Then, instead of performing an individual gene (variable) analysis, we are assuming the Pearson's linear correlation coefficient as the epistatic relation of the variables in our datasets (see Chapter 5). Such statistical measure makes easy to assess the interdependence among variables [56, 63].

The goal consists in performing an epistasis analysis by using the Pearson's linear correlation coefficient (ρ) and avoiding the schemata disruption. Avoiding the schemata disruption is crucial to prevent the problem decomposition, especially when dealing with high data dimensionality and interdependent subproblems. Therefore, we are proposing two different versions of EbFSA (EbFSA_v1 and EbFSA_v2). Our both deterministic approaches are based on the Compact Genetic Algorithm (Algorithm 3.1) and concept of epistasis (Section 3.5.3). Instead of creating a population of individuals and applying a stochastic search, EbFSA uses only one chromosome to represent and deterministically select the most informative variables to compose the calibration model.

4.6.1 EbFSA_v1

In the first version of EbFSA, the chromosome is initially created by selecting the most linear independent variables (columns) in the $n \times k$ matrix $\mathbf{X}_{cal}$. To this end, the

Pearson's linear correlation coefficients matrix $\mathbf{R}_{k\times k}$ is obtained from $\mathbf{X}_{cal}$. For instance, if a variable with index j is (near) orthogonal to another variable with index i ($\mathbf{R}_{j,i} \approx 0$), then both variables (genes) are set from 0 to 1 in the chromosome.

Soon after, matrix $\mathbf{R}_{k\times k}$ is analyzed again and the most correlated variables are selected. Thereby, two correlated variables ($\mathbf{R}_{j,i} \approx 1$) remain together in the same subset avoiding possible disruptions. However, the model prediction error (RMSEP) is calculated at every variables insertion in this stage. Thus, if two dependent variables do not collaborate to reduce RMSEP, then both variables are discarded. Algorithm 4.5 shows the pseudocode for EbFSA_v1:

Algorithm 4.5: Proposed EbFSA_v1.

```
1:  Obtain matrix R_{k×k} from matrix X_cal by using (for example) corrcoef Matlab© built-in function
2:  Generate an unique individual as a null row vector v_{1×k} = {0, 0, ..., 0}
3:  for i = 1 : k
4:     for j = (i + 1) : k
5:        if R_{j,i} < 0.0001
6:           v_i = 1
7:           v_j = 1
8:     end for j
9:  end for i
10: Calculate and assess RMSEP according to Equation (2-3)
11: for i = 1 : k
12:    for j = (i + 1) : k
13:       if R_{j,i} > 0.9999
14:          v_i = 1
15:          v_j = 1
16:          Discard both variables if RMSEP was not reduced
17:    end for j
18: end for i
19: Plot all selected variables
```

In Algorithm 4.5, as matrix $\mathbf{R}_{k\times k}$ is symmetric, only its lower triangular part is traversed to select the appropriate variables. In both stages, a precision of 10^{-4} (0.0001 and 0.9999) is empirically used. Note that a greater precision may force the algorithm to select a more reduced number of variables, which could imply in reducing the model predictive ability (*i.e.*, increasing RMSEP). On the other hand, a lower precision may cause the algorithm to select a larger number of variables, and this could also increase the model prediction error.

4.6.2 EbFSA_v2

In the second version of EbFSA, we use a different strategy. Instead of analyzing matrix $\mathbf{R}_{k\times k}$ for every variable, it is obtained the sum of each column of matrix $\mathbf{R}_{k\times k}$. The value of each column sum is ordered, and the row vector $\mathbf{index}_{1\times k}$ contains the indexes of the sorted values. Then, one variable is selected at a time according to the index stored in $\mathbf{index}_{1\times k}$. If its fitness value (RMSEP) does not improve the model predictive ability, the variable is discarded. At the end, the model contains only the most informative variables. Algorithm 4.6 shows the pseudocode for EbFSA_v2.

Algorithm 4.6: Proposed EbFSA_v2.

1: Obtain matrix $\mathbf{R}_{k\times k}$ from matrix $\mathbf{X}_{cal}$ by using (for example) *corrcoef Matlab*© built-in function
2: Generate an unique individual as a null row vector $\mathbf{v}_{1\times k} = \{0, 0, ..., 0\}$
3: Obtain the columns sum values of matrix $\mathbf{R}_{k\times k}$ in vector $\mathbf{sum}_{1\times k}$
4: Sort the values of vector $\mathbf{sum}_{1\times k}$ and store the indexes of the sorted values in vector $\mathbf{index}_{1\times k}$
5: **for** $i = 1 : k$
6: $\quad \mathbf{v}_{\mathbf{index}_i} = 1$
7: $\quad$ Check RMSEP according to Equation (2-3)
8: $\quad$ Discard this variable if RMSEP was not reduced
9: **end for** i
10: Plot all selected variables

The strategy used in Algorithm 4.6 is based on columns of matrix $\mathbf{R}_{k\times k}$. Each column index in matrix $\mathbf{R}_{k\times k}$ represents the respective column (variable) index in matrix $\mathbf{X}_{cal}$. If the sum of a column values is low, it means the variable is minimally correlated to the other variables. On the other hand, if the sum of a column values is high, then such variable is strongly correlated to the others.

It is clear to see EbFSA_v2 has a more improved strategy than EbFSA_v1. In the first version, two selected variables can be discarded together if both do not reduce RMSEP. This issue may ignore some relevant variable previously selected. Differently, EbFSA_v2 is able to select informative variables based on epistatic relation through the sum of the Pearson's linear correlation coefficients of each column in matrix $\mathbf{R}_{k\times k}$.

4.7 Summary

When the problem can not be properly decomposed, standard GAs with binary representation tend to lead to an undesirable performance. In this context, decision variables from spectroscopic data in multivariate calibration models usually contain

strong multicollinearity and an attempt of decomposition by crossover can directly affect the final result.

In this work, two hypotheses are proposed. Hypothesis 1 states variable selection procedure in multivariate calibration can be considered as a non-completely decomposable problem. Three numerical examples provide significant evidences about the viability of such hypothesis and demonstrate Equation (4-2) can not be satisfied. Our first hypothesis is already published by Paula *et al.* [83].

Building block is a low-order and short defining-length schema with an above-average fitness. It is able to generate the best subjects in a GA population. In this sense, Hypothesis 2 claims recombination operators tend to cause schemata disruption. One proposition and three additional numerical examples show that, although not necessarily there exists building blocks formation in spectral data from multivariate calibration due to high dimensionality, the disruption of schemata directly affects the non-decomposability assumption in the first hypothesis. Our second hypothesis is already published by Paula *et al.* [84].

It is noteworthy the numerical examples for our both hypotheses are relatively simple and lack mathematical proofs to demonstrate their generalization. This issue extrapolate the goal of this work, which consists (not only) in showing their veracity and applicability in our large datasets.

We propose a GA implementation to provide additional evidences about the hypotheses feasibility. We also present one heuristic for initial individuals generation; another heuristic strategy for possible schemata identification; and one simple local search operator to reduce schemata disruption. They are all empirically implemented in order to improve the standard GA implementation.

Additionally, two different versions of a novel approach for variable selection are presented: Epistasis-based Feature Selection Algorithm (EbFSA). EbFSA avoids schemata disruption by preventing the problem decomposition. It uses only one individual and the Pearson's linear correlation coefficient as an epistatic relation in order to select the best variables to compose the calibration model.

CHAPTER 5

Experimental

This work uses two different datasets. Dataset 1 consists on wheat samples, and dataset 2 corresponds to gas mixtures. Section 5.1 and Section 5.2 describe these both datasets, respectively.

5.1 Dataset 1

Dataset 1 employed in this work consists in 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.* [85] and Soares *et al.* [101]. The dataset for the multivariate calibration study consists on 690 Near-Infrared Reflectance (NIR) spectra of whole-kernel wheat samples, which were used as shoot-out data in the 2008 International Diffuse Reflectance Conference [27].

Protein concentration in the analyzed 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. In this work, the employed NIR was in range of 1100-2500 nm. In order to remove undesirable features, the first derivative spectra were calculated by using the Savitzky-Golay filter with a second-order polynomial and an eleven point window [85, 87, 104].

Reference values of the protein concentration in the wheat samples were determined in laboratory by Kjeldahl method [11]. Such method causes the organic substances destruction with concentrated sulfuric acid in the presence of a catalyst and by the action of heat, and subsequent distillation of nitrogen from the sample. On the other hand, the use of indirect instrumental techniques such as NIR and mathematical models (MLR) allow protein level to be determined without destroying the sample.

Kennard and Stone [57] algorithm was applied to the resulting spectra to divide the samples into three sets: calibration, validation and prediction. Calibration set contains 389 samples with 690 variables each, and it was used to calculate the regression coefficients. Validation and prediction sets contains 193 samples both. Validation set was

employed to guide the variable selection in GA. Prediction set was only employed in the final performance assessment of the resulting regression model.

Figure 5.1 plots the NIR spectra of dataset 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 disturbance (interference) implying in considerable rapprochement (dependency) among variables from the spectra [23, 39, 66]. This issue means that in most of wavelength regions there may be a relatively large number of correlated variables.

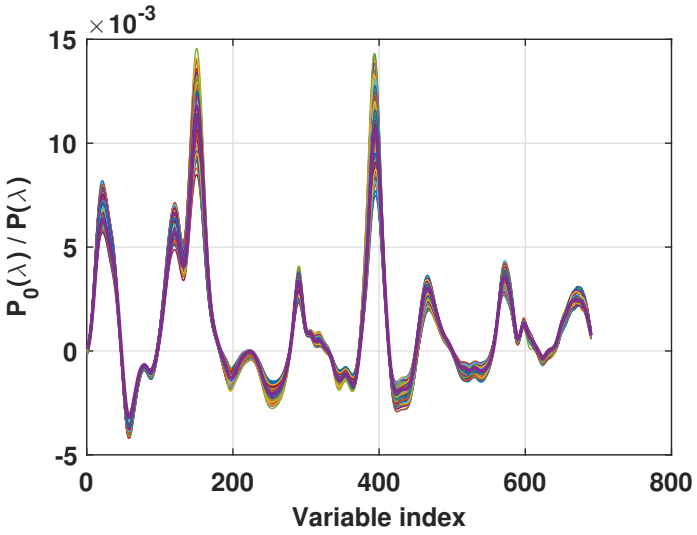

Figure 5.1: *NIR spectra of dataset 1 [83].*

5.2 Dataset 2

In this work, we also used one petrochemical dataset composed of samples generated on a laboratory scale in transmittance. Such dataset was used as shoot-out data in the International Diffuse Reflectance Conference (IDRC) 2014 (www.idrc-chambersburg.org) [55]. The sample in this dataset corresponds to gas mixtures in the gas phase measured in transmittance. It was collected in laboratory under various conditions of pressure and temperature, and the data correspond to NIR spectra.

Four participants competed in IDRC 2014. Each participant performed a pre-processing to apply their own techniques in regions which they consider to have greater climate influences. However, in our approach we did not perform any pre-processing to select specific regions of spectra in order to obtain better outcomes [106].

Dataset spectra was divided by IDRC organization into three sets: calibration, validation and test (or prediction). Each one contains 144, 60 and 36 samples with 3999 features (variables), respectively. Calibration samples were used to obtain the regression

model. Validation set was employed to guide the selection of variables. Test set was used in the final performance assessment of the resulting MLR model.

Figure 5.2 plots the NIR spectra of dataset 2. The spectra also come from the calibration set. As in Figure 5.1, it is possible to note a significant approximation between the spectral regions which indicates a considerable correlation in the spectral regions.

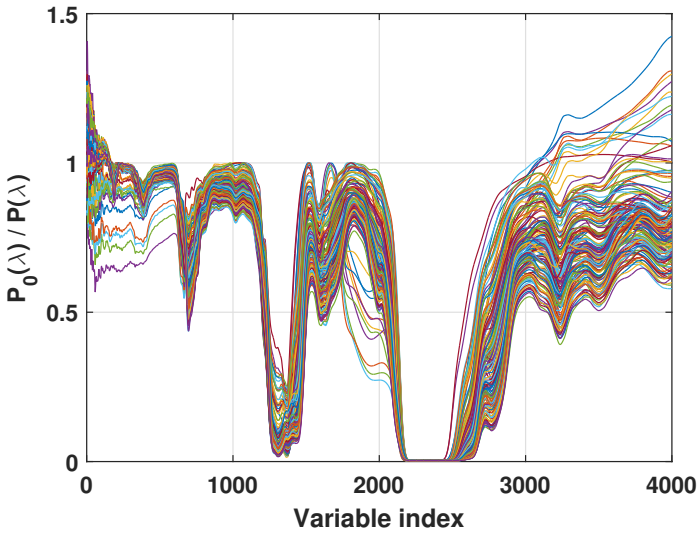

Figure 5.2: *NIR spectra of dataset 2.*

5.3 Computational Platform

All tests were performed in a desktop personal computer with a Intel core i7 2600 (3.4 GHz) processor and 8 GB of RAM memory. *Matlab*© R2015a software has been used for simulations and data processing. The built-in function *pinv*() was applied to perform the matrix inverse calculation involved in Equation (2-2). Finally, it is noteworthy that all outcomes were obtained by averaging 30 trials each.

5.4 Summary

This doctoral thesis uses two datasets as benchmark in order to generate and obtain outcomes. Dataset 1 consists in whole-wheat grain samples. Such samples were used as shoot-out data in the IDRC 2008 [27] and have been used by several works from literature [64, 85, 87, 101, 104]. Dataset 2 consists in gas mixtures samples. It was used as shoot-out data in the IDRC 2014 [55], which contained four participants and participant 3 was the champion. The tests were performed in a specific computer and a scientific software have been used to simulate and process the datasets.

CHAPTER 6

Results and Discussion

This chapter presents all results obtained by using our proposed algorithms. Initially, Section 6.1 presents a linear correlation analysis in dataset 1 and dataset 2 in order to reinforce the non-decomposability assumption approached by the first hypothesis (Section 4.1.2) due to the multicollinearity problem.

In order to enhance the investigation performed by the second hypothesis (Section 4.2.1), Section 6.2 presents a schemata disruption analysis using two different approaches: *i*) crossover operator; and *ii*) local search. The first one utilizes crossover operator to split the individuals in the GA population in order to form new offspring. The last one is based on a simple local search, which aims to reduce schemata disruption usually caused by recombination operators.

Section 6.3 provides the outcomes from both datasets by EbFSA_v1 and briefly discuss the correlation between neighbor variables in spectral regions. Section 6.4 presents the outcomes from dataset 1 and dataset 2 obtained by EbFSA_v2. Finally, Section 6.5 shows a comparison between our proposed algorithms and different techniques from literature.

6.1 Non-Decomposability Assumption

6.1.1 Linear correlation analysis in dataset 1

Figure 6.1 [1] shows a hot color map representing the correlation among all variables from Figure 5.1 (dataset 1):

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

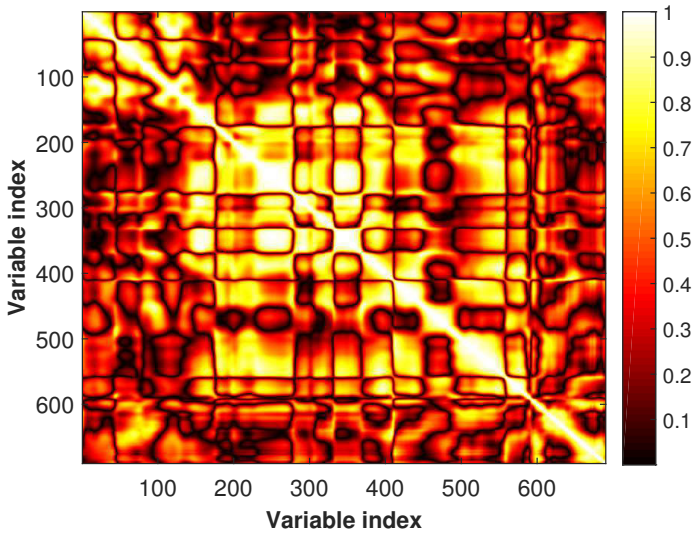

Figure 6.1: *Linear correlation analysis among all variables from dataset 1 [83].*

The *corrcoef Matlab*© built-in function yields a symmetric matrix $\mathbf{R}_{690\times690}$ calculated from the input matrix $\mathbf{X}_{389\times690}$ (dataset 1) whose rows are observations (samples) and columns are features (variables). This function is calculated by Pearson's linear correlation coefficient (see Section 2.4.3). 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.

One can notice in Figure 6.1 that in fact there is considerable correlation among most of variables. For example, variable 320 and variable 480 yield Pearson's linear correlation coefficient $\rho_{320,480} = 0.0172$ indicating they are minimally correlated and could be possibly selected to contribute for the increasing of model predictive ability. On the other hand, variables 646 and 647 yield $\rho_{646,647} = 0.9940$, which indicates they are almost totally correlated and both contribute to increase the multicollinearity in the model. It is important to note that by selecting the most orthogonal variables becomes possible to reduce the multicollinearity and obtain an adequate accurate model [2] [23]. However, not all orthogonal variables provide relevant information about the property of interest [56].

Through Figure 6.2 it is possible to graphically notice that indeed variables 646 and 647 from Figure 6.1 are strictly correlated over the 389 samples:

[2] In the context of this work, an accurate model implies in a model with a considerable predictive ability.

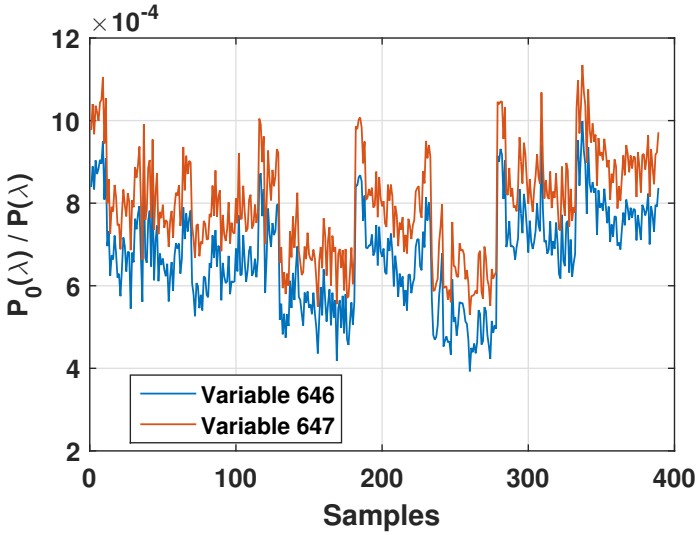

Figure 6.2: *Linear correlation analysis between variables 646 and 647 from Figure 6.1.*

Figure 6.3 shows variable 593 is the one who presents the lowest level of correlation between any other variable from dataset 1. It is possible to perceive the predominant dark color in the chart, which indicates that, in fact, variable 593 is linearly independent of most other variables ($\rho_{593,i} \approx 0$, where $1 \leq i \leq 690$ and $i \neq 593$). The clear dash line highlights variable 593 is directly correlated with itself ($\rho_{593,593} = 1$).

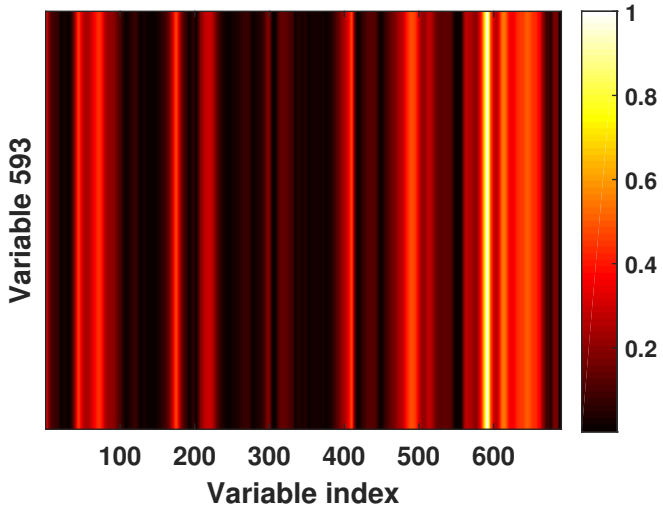

Figure 6.3: *Linear correlation analysis between variable 593 and the others from dataset 1.*

Similarly, Figure 6.4 demonstrates variable 548 is the one who presents the highest level of correlation among the others from dataset 1. The predominant bright color indicates its significant linear correlation ($\rho_{548,i} \approx 1$, where $1 \leq i \leq 690$ except a few variables). Precisely because variable 548 is strongly correlated with the majority, its

presence in the model may become important since it can carry relevant information from the others [56].

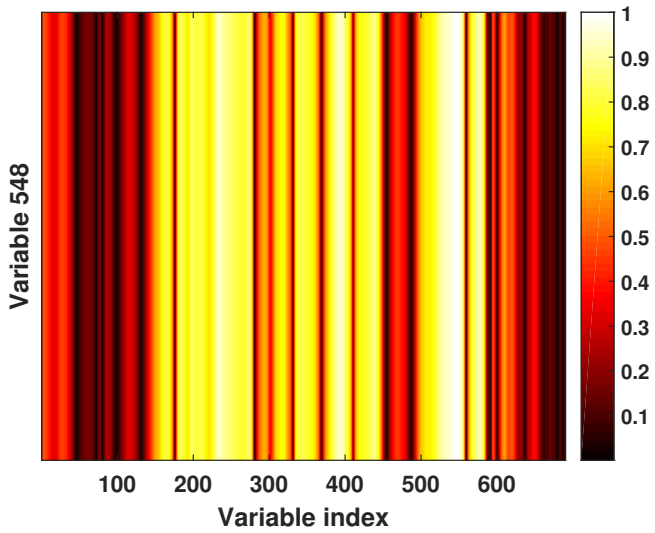

Figure 6.4: *Linear correlation analysis between variable 548 and the others from dataset 1.*

6.1.2 Linear correlation analysis in dataset 2

Considering Figure 5.2 in the last chapter, Figure 6.5 shows a hot color map representing the linear correlation among all variables from dataset 2 (see Section 5.2):

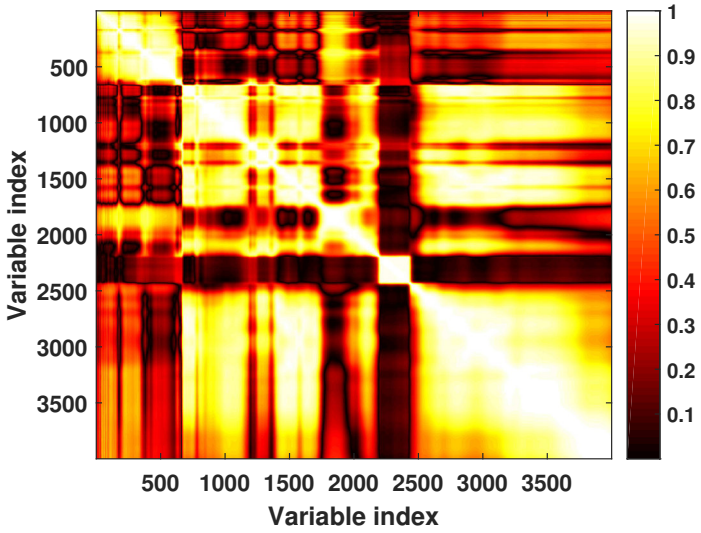

Figure 6.5: *Linear correlation analysis among all variables from dataset2.*

As in Figure 6.1, there is considerable correlation among most of variables in dataset 2. For instance, the spectral region between variables 2500 and 3999 is the most

bright in the chart, which indicates it contains a large number of correlated variables (high level of multicollinearity).

Figure 6.6 shows variable 2 is the most independent in relation to all other variables from dataset 2:

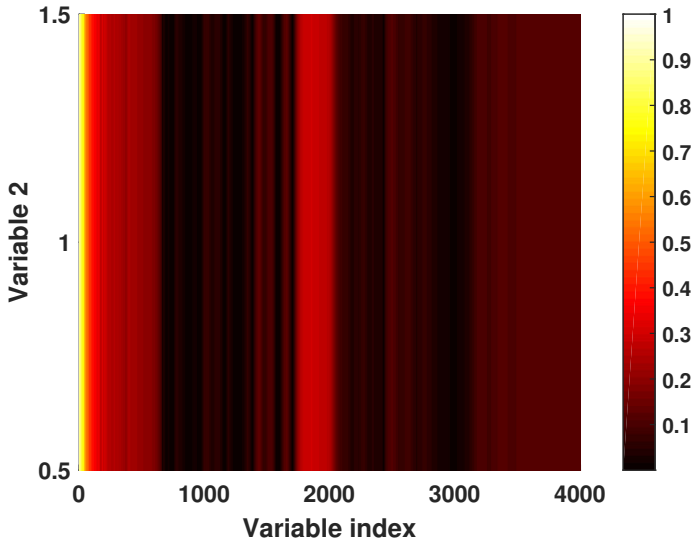

Figure 6.6: *Linear correlation analysis between variable 2 and the others from dataset 2.*

On the other hand, Figure 6.7 shows variable 3202 is the most correlated to all other variables from dataset 2:

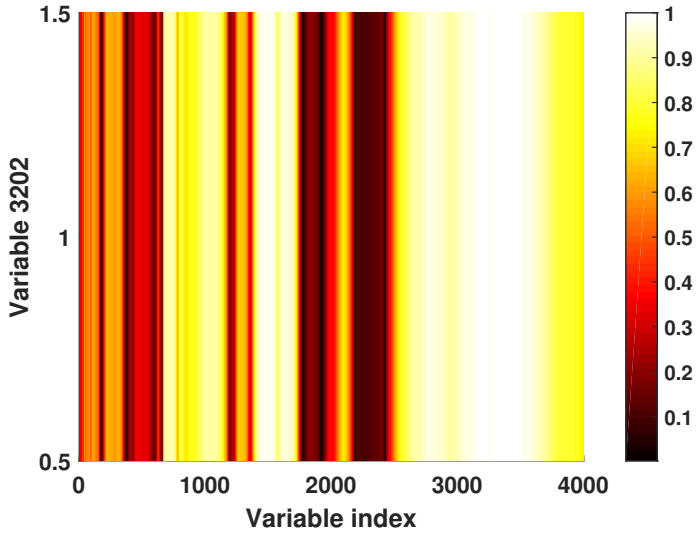

Figure 6.7: *Linear correlation analysis between variable 3202 and the others from dataset 2.*

Finally, it is noteworthy that Figures 5.1, 5.2, 6.1, 6.2, 6.3, 6.4, 6.5, 6.6 and 6.7 provide additional evidences about the non-decomposability assumption for the

variable selection procedure in multivariate calibration. Therefore, they strengthen our first hypothesis (see Section 4.1.2).

6.2 Analysis of Possible Schemata Disruption

6.2.1 Using crossover operator in dataset 1

Figure 6.8 shows an analysis of schemata disruption using the proposed GA implementation with crossover operator (see Algorithm 4.1):

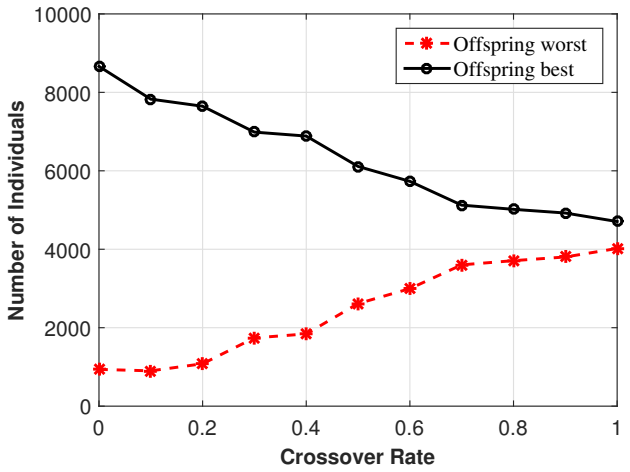

Figure 6.8: *Schemata disruption analysis using dataset 1 [84].*

It is possible to observe the offspring from worst individuals tends to increase while the offspring from best individuals [3] tends to decrease. It indicates an evidence that the possibility of schemata disruption raises as crossover rate increases, which directly implies in our fist hypothesis (Section 4.1.2).

On the one hand, if selective pressure [4] is high, an uniform population can be quickly obtained, and this will lead to the loss of diversity. On the other hand, a reduced disruption factor can decrease the crossover probability, and this will lower the number of schemata exchange. Thus, we claim this may occur due to the fact that a schema can be lost a few times in many possible cross sites and its disruption factor comes down to a specific value [30, 41, 47]. Moreover, this issue directly affects the non-decomposability assumption of the variable selection procedure in multivariate calibration (see Section 4.2.1).

[3] It is important to note that in the context of Figure 6.8, the best individuals are those who select the smallest number of variables.

[4] Selective pressure is the tendency to select only the best individuals of the current generation to propagate to the next.

Figure 6.9 shows the obtained-prediction error values as well as schema fitness using crossover operator:

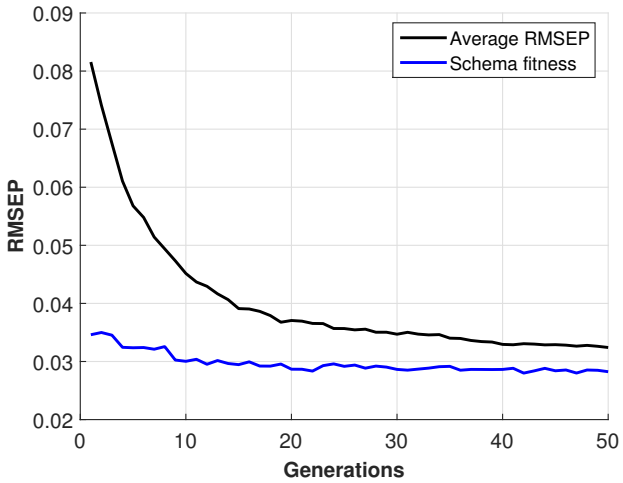

Figure 6.9: *Schemata disruption analysis using dataset 1 and a GA population with 500 individuals.*

In this simulation, we do not use selective pressure neither elitism [5]. Instead, each parent participates in a single crossover and the parents are always replaced by offspring [6]. Consequently, the average RMSEP and schema fitness curves may vary over generations.

Schema fitness was obtained by using the strategy presented in Algorithm 4.3 with 50% of the best individuals. It is clear to notice in Figure 6.9 the average RMSEP values tend to decrease over generations. Finally, schema fitness provides a considerably-reduced initial RMSEP value. As described in Section 3.2, if schemata are short and has low-order, the number of desirable schemata tends to increase. Otherwise, crossover operator may frequently disrupt high-order or large schemata.

Figure 6.10 presents the same comparison in the previous figure. The difference here is the use of the heuristic presented in Algorithm 4.2 (initial solutions generation).

[5] Elitism means that at least one changeless copy of the best solution of current generation is transferred to the new population so that the best solution can survive to successive generations.

[6] The application of operators without any selective pressure does not modify the population statistics properties [74].

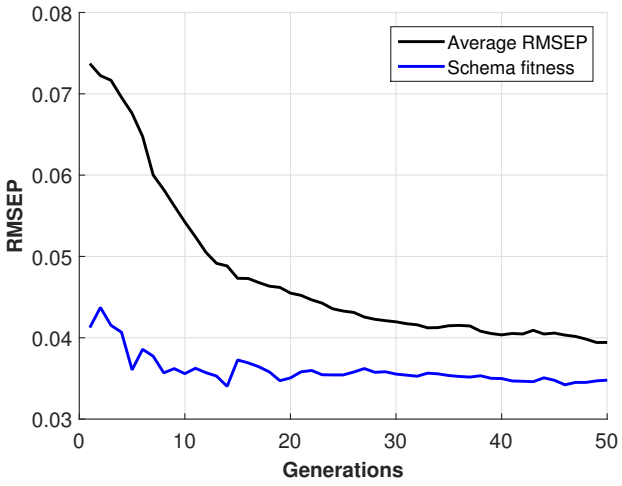

Figure 6.10: *Schemata disruption analysis using dataset 1 and the initial solutions heuristic with 500 individuals.*

One can observe the use of such strategy directly implies in a more reduced initial RMSEP value due to the fact that (near) orthogonal variables are initially selected. However, it is also possible to observe an even more significant variation in schemata fitness curve. Although providing RMSEP values below the average, this issue is a relevant evidence that the best schemata may have been broken at some generations, which implies in the abrupt curve changes.

6.2.2 Using crossover operator in dataset 2

Figure 6.11 shows the average RMSEP and schema fitness comparison using dataset 2:

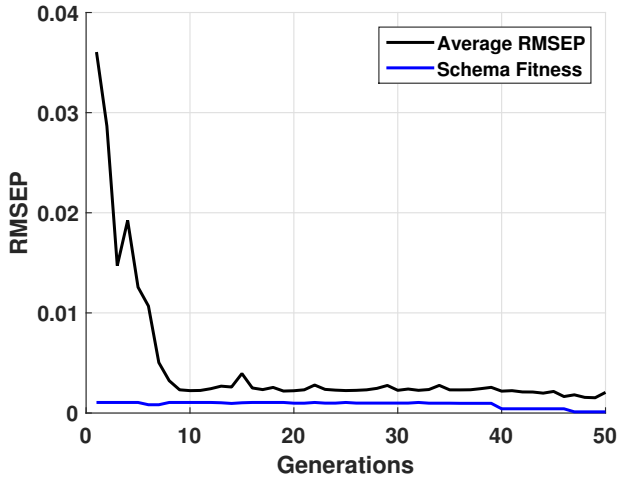

Figure 6.11: *Schemata disruption analysis using dataset 2 and a GA population with 500 individuals.*

The average RMSEP curve stabilizes at (approximately) the 10th generation while the schema fitness curve remains almost constant. In this case, as we do not use selective pressure neither elitism, the best individuals in the population are not preserved. Consequently, the RMSEP value may change abruptly, which could explain some variations in the average RMSEP curve. Moreover, considering the strategy used in Algorithm 4.3, we believe the best schemata (*i.e.*, the chromosomes of the best individuals) were not able to generate the 50% of the best individuals in the population, which corroborates the curve constancy.

Finally, Figure 6.12 shows the same comparison using the heuristic presented in Algorithm 4.2. It is possible to notice a similar behavior of both curves in comparison with Figure 6.11. Although providing a more reduced initial average RMSEP value, such issue indicates the use of the heuristic (Algorithm 4.2) in this case has not significantly improved the outcomes (see Table 6.5 for more details).

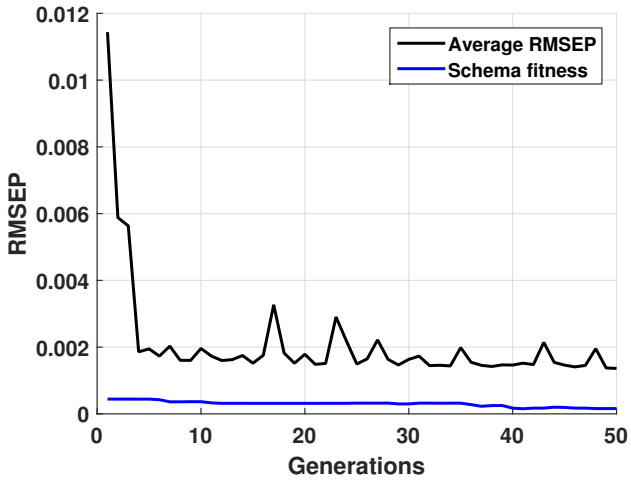

Figure 6.12: *Schemata disruption analysis using dataset 2 and the initial solutions heuristic with 500 individuals.*

6.2.3 Using local search operator in dataset 1

Figure 6.13 shows the obtained results by using the proposed local search-based operator (see Algorithm 4.4):

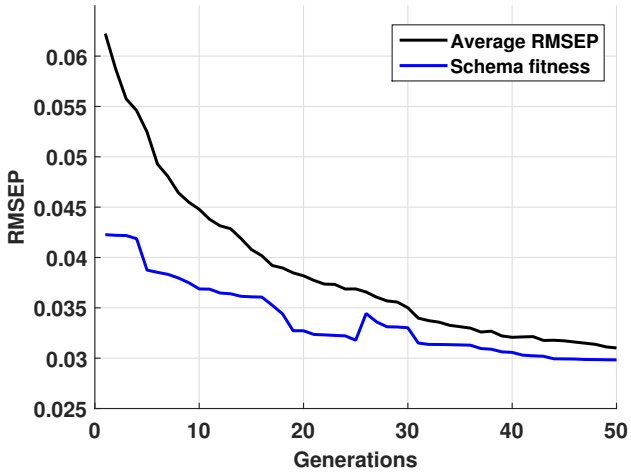

Figure 6.13: *Schemata disruption analysis using dataset 1 with initial solutions heuristic, local search operator and 500 individuals.*

It is possible to notice a significant initial RMSEP reduction as well as a more reduced schema-fitness curve. Such chart demonstrates the use of a simple local search operator is able to provide better outcomes when compared with crossover operator.

6.2.4 Using local search operator in dataset 2

Figure 6.14 shows the same comparison in Figure 6.12 but using local search instead of crossover operator. In this case, the local search operator (besides also converging quickly) provided soft curves indicating a significant reduction in the schemata disruption.

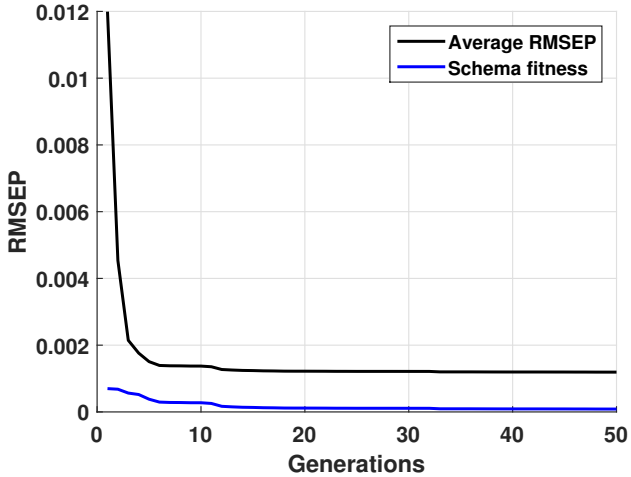

Figure 6.14: *Schemata disruption analysis using dataset 2 with initial solutions heuristic, local search operator and 500 individuals.*

The local search improves the variable neighborhood exploration and tends to reduce the schemata disruption [94]. However, even reducing the schemata disruption, the local search employed in Algorithm 4.4 is simple and can cause it at some generations. For example, the RMSEP value has been steadily reduced until the 25th generation in Figure 6.13. At that point, its value was sharply increased which indicates the best schema could be possibly disrupted.

Finally, it is important to highlight that Figures 6.8, 6.9, 6.10, 6.11, 6.12, 6.13 and 6.14 provide additional and relevant evidences for the schemata disruption investigation approached in Section 4.2.1. Then, they all fortify our second hypothesis.

6.3 Results for Epistasis-based FSA version 1

Table 6.1 summarizes all obtained outcomes by using the proposed EbFSA_v1:

Table 6.1: *EbFSA_v1 outcomes.*

Dataset 1	**EbFSA_v1**
RMSEP	0.0624
MAPE	1.46%
PRESS	12.04
Number of variables	57
Dataset 2	**EbFSA_v1**
RMSEP	0.0050
MAPE	3.78%
PRESS	0.0009
Number of variables	249

It is possible to see EbFSA_v1 yielded significantly reduced model prediction error values when applied into dataset 2. On the other hand, it selected a smaller number of variables from dataset 1. Since dataset 2 has a considerably larger number of variables (almost 4000), we believe the algorithm tends to select more informative variables to compose the calibration model. However, Section 6.4 shows this does not happen when using EbFSA_v2.

Figures 6.15 and 6.16 show the variables (red dots) in the spectral regions selected by EbFSA_v1 from dataset 1 and dataset 2, respectively:

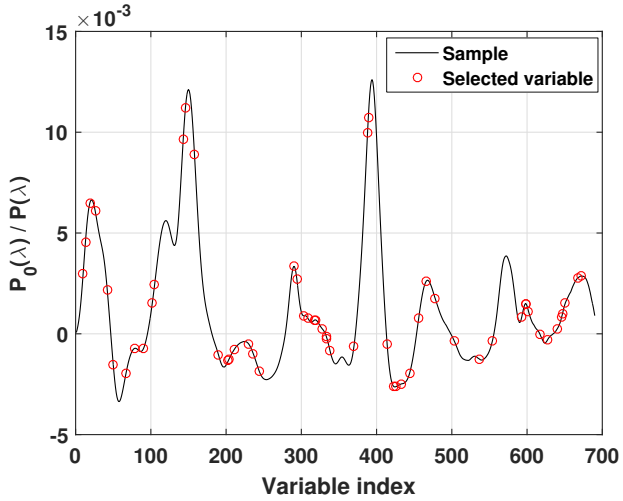

Figure 6.15: *Selected variables by EbFSA_v1 from dataset 1.*

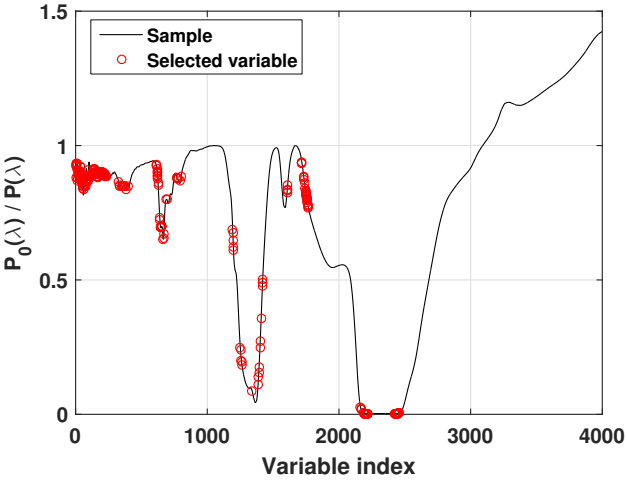

Figure 6.16: *Selected variables by EbFSA_v1 from dataset 2.*

Such variables were selected based on the Pearson's linear correlation coefficients in matrix $\mathbf{R}_{k\times k}$, which is obtained from matrix $\mathbf{X}_{n\times k}$ ($\mathbf{X}_{389\times 690}$ for dataset 1, and $\mathbf{X}_{144\times 3999}$ for dataset 2). Those coefficients were considered as the epistatic relation among variables (see Section 3.5.3). For instance, correlation coefficient value close to zero ($\rho_{i,j} < 0.0001$) indicates two variables are (near) linearly independent and both could be selected [7]. Correlation coefficient values close to 1 ($\rho_{i,j} > 0.9999$) indicate two variables are strictly correlated, and they are selected to compose the model only if the model prediction error is reduced (they are discarded otherwise). For more details, Algorithm 4.5 must be checked. It is important to note the analysis of variables spread over different spectral regions is beyond the scope of this work.

[7] As discussed in Chapter 2, independent variables usually provide more significant information about the problem [39].

Below are the index of all (fifty-seven) variables in Figure 6.15:

10 13 19 26 42 50 67 78 90 102 105 143 146 158 190 203 204 211 230 236 244 290 294 303 309 318 319 328 333 334 338 370 388 389 414 423 425 433 444 456 466 478 503 537 554 593 598 599 602 617 627 640 646 647 651 668 672.

These are the most informative variables selected by EbFSA_v1 from dataset 1. Most of them are linearly independent (orthogonal) of each other. However, one single variable is not necessarily orthogonal to all other variables. For example, variables 10 and 294 have $\rho_{10,294} = 0.0001$ (close to zero), which indicates they are (near) linearly independent. On the other hand, variables 10 and 13 have $\rho_{10,13} = 0.9999$ (close to 1) indicating they are strongly correlated. Although variables 10 and 13 are correlated, they are (near) orthogonal to at least one other selected variable and, precisely because they are correlated, both together in the same subset tend to collaborate to reduce the model prediction error. As previously discussed, the presence of one without the other usually implies in loss of performance since a dependent variable carries information from the other one to which it linearly depends [56].

Figure 6.17 shows a zoom in between variable indexes 600 and 690 in Figure 6.15. It is possible to see in such spectral region the selected variables 602, 617, 627, 640, 646, 647, 651, 668 and 672 (the nine last variables in Figure 6.15). This chart demonstrates those selected variables are independent to at least one other selected variable spread over the sample, but two or more neighboring variables can be correlated. Therefore, some variables are correlated to their neighboring variables [8].

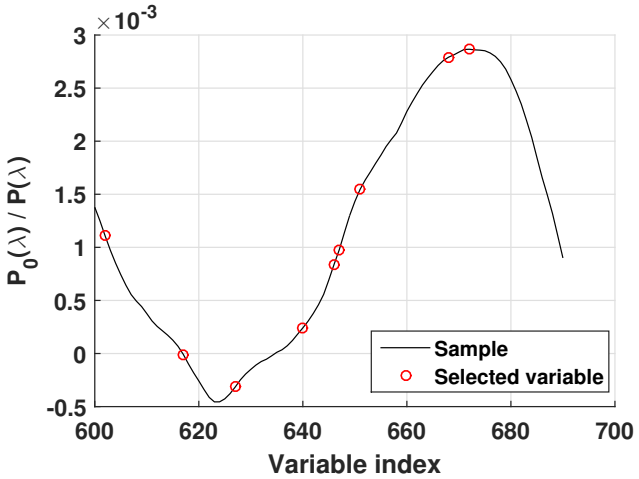

Figure 6.17: *Zoom in the spectral region between indexes 600 and 690 in Figure 6.15.*

[8] Figure 6.2 graphically demonstrates variable 646 is strongly correlated with variable 647, which is its neighbor.

According to Atkins [7], all the variables which make up the same spectroscopic band (spectral region) may be highly correlated to each other as they result from the same physical process. That is, a band on the spectrum represents a spectroscopic transition which is associated with the molecular structure. Thus, all variables of this same band represent the same phenomenon. Because of the spectroscopic effect of band enlargement, it is possible to check bands instead of a vertical line representing the transition. Then, this enlargement can cause the variables close to the one of greater intensity to be highly correlated because they result from the same observed phenomena.

Spectra may also show band overlays, which makes a band appear as irregular [7]. In this case, although it looks like a same band, some variables result from different phenomena with different variabilities and correlations between the variables of the (apparent) same band. Therefore, it is to be expected the nearest variables are the most correlated among themselves even when they make up the (probable) same band [7, 8]. Finally, it is important to emphasize it is beyond the scope of this work to investigate the reason of why such phenomena happens.

6.4 Results for Epistasis-based FSA version 2

Table 6.2 summarizes all obtained outcomes by using the proposed EbFSA_v2 and compares them with EbFSA_v1:

Table 6.2: *Outcomes comparison between EbFSA_v1 and EbFSA_v2.*

Dataset 1	**EbFSA_v1**	**EbFSA_v2**
RMSEP	0.0624	0.0792
MAPE	1.46%	1.58%
PRESS	12.04	15.28
Number of variables	57	131
Dataset 2	**EbFSA_v1**	**EbFSA_v2**
RMSEP	0.0050	0.0018
MAPE	3.78%	6.56%
PRESS	0.0009	0.0001
Number of variables	249	7

It is possible to see EbFSA_v1 is better than EbFSA_v2 regarding dataset 1. Nevertheless, EbFSA_v2 is able to overcome EbFSA_v1 by selecting a considerably reduced number of variables from dataset 2 and obtaining a more reduced RMSEP value.

Figures 6.18 and 6.19 illustrate all variables in the spectral regions selected by EbFSA_v2 from both datasets. Such as in EbFSA_v1, variables were picked based on the Pearson's linear correlation coefficients (see Algorithm 4.6).

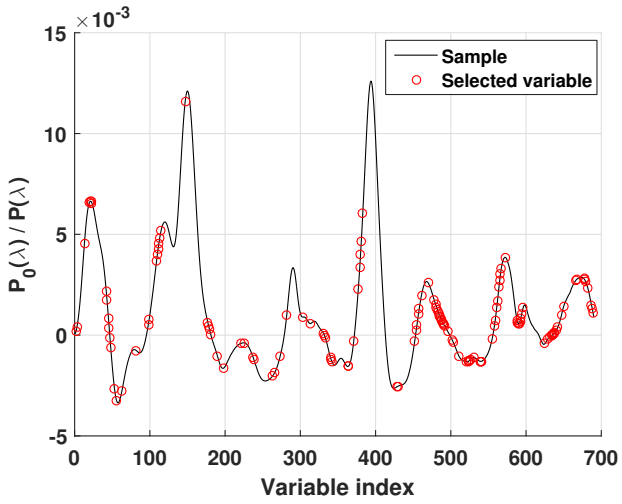

Figure 6.18: *Selected variables by EbFSA_v2 from dataset 1.*

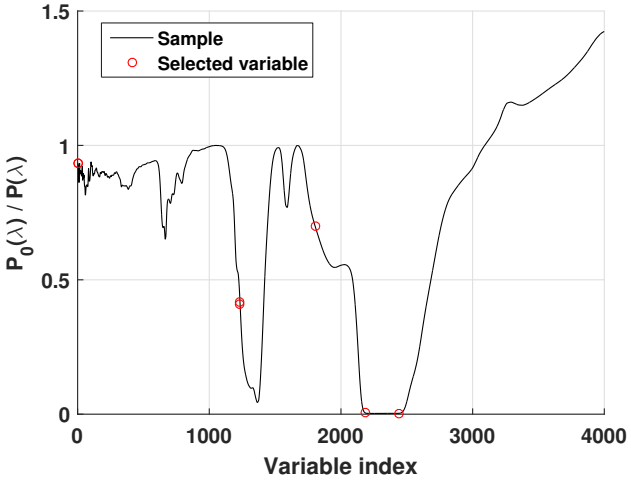

Figure 6.19: *Selected variables by EbFSA_v2 from dataset 2.*

Considering Figure 6.19, the index of all (seven) selected variables from dataset 2 are: 1 2 1233 1234 1803 2187 2441. Table 6.3 shows the Pearson's linear correlation coefficient values between each one [9]:

[9] As matrix **R** is symmetric (see Example 3 in Section 4.1.3), only its superior triangular part is showed.

Table 6.3: *Pearson's linear correlation coefficient between each selected variable from dataset 2 by EbFSA_v2.*

Variable	1	2	1233	1234	1803	2187	2441
1	1	0.9863	0.0496	0.0399	0.2587	0.0527	0.0158
2		1	0.0233	0.0222	0.2847	0.0361	-0.0199
1233			1	0.9999	0.3932	0.8844	0.3644
1234				1	0.3912	0.8845	0.3666
1803					1	0.4360	-0.0736
2187						1	0.3060
2441							1

These are the most informative variables selected by EbFSA_v2 (Algorithm 4.6). As shown in Figure 6.6, variable 2 is the most independent one of the others. However, Table 6.3 shows variable 2 is strongly correlated with variable 1 ($\rho_{1,2}$ = 0.9863). Consequently, their presence in the model becomes indispensable to obtain a more reduced RMSEP value since one is linearly dependent of the other.

Additionally, Figure 6.20 shows variables 1233 and 1234 are totally linearly dependent ($\rho_{1233,1234} = 0.9999$). Indeed, as discussed earlier some neighbor variables tend to be strictly correlated. Therefore, one can graphically observe they are closely linearly spaced and both together contribute to reduce the model prediction error.

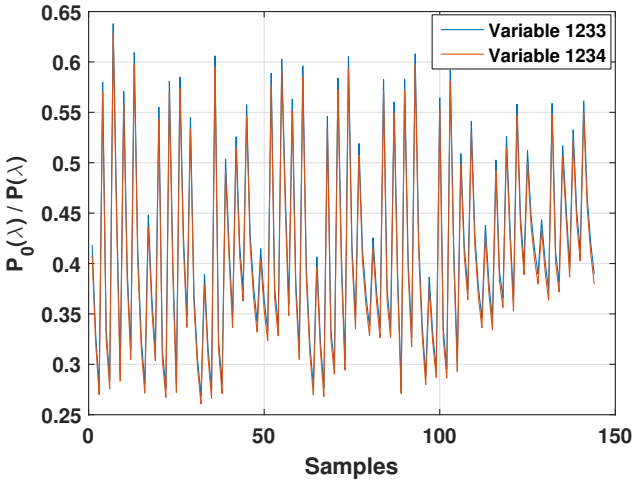

Figure 6.20: *Linear correlation analysis between variable 1233 and 1234 from dataset 2.*

Finally, Figures 6.21 and 6.22 show us the RMSEP values during the variable selection process from dataset 1 and dataset 2 obtained by EbFSA_v2, respectively:

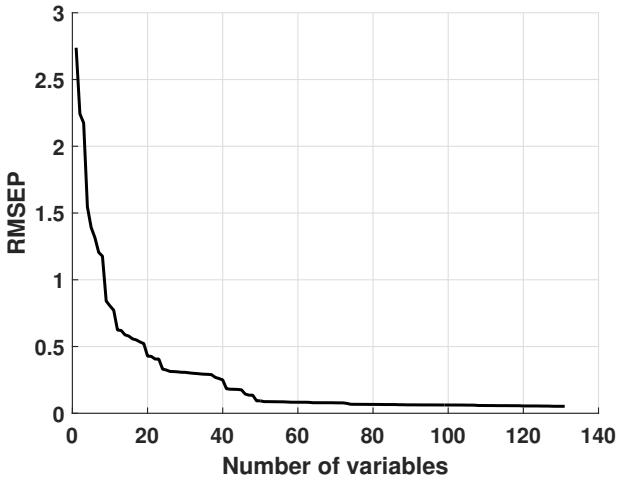

Figure 6.21: *RMSEP values during variable selection from dataset 1 by EbFSA_v2.*

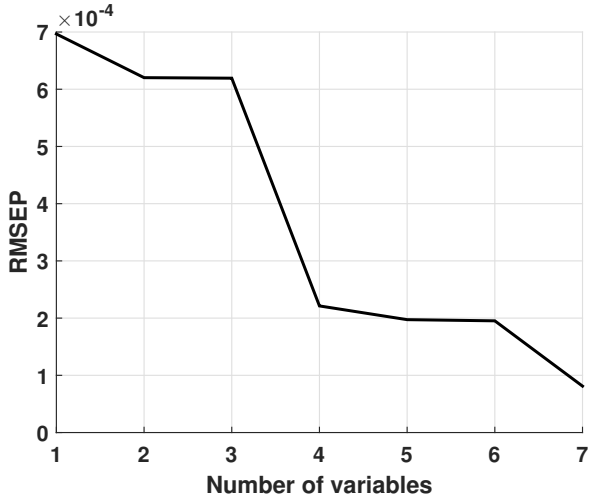

Figure 6.22: *RMSEP values during variable selection from dataset 2 by EbFSA_v2.*

One can notice the prediction error reduction as the variables are inserted into the calibration model. However, it is important to note the values in the charts does not necessarily match the final RMSEP value (see Table 6.2) because the prediction set ($\mathbf{X}_{pred}$) is used at the final evaluation while the validation set ($\mathbf{X}_{val}$) is used during the variable selection process. See Chapter 5 for more details.

6.5 Literature Comparison

Table 6.4 presents a comparison between our proposal and some different algorithms from literature applied into dataset 1:

Table 6.4: *Outcomes comparison between different algorithms using dataset 1.*

Algorithm	RMSEP	MAPE	PRESS	Num. of variables
PLS [85]	0.21	1.50%	10.09	15
SPA-MLR [85]	0.20	1.43%	9.95	13
BVS [85]	0.15	1.07%	6.96	29
FA-MLR [85]	0.09	0.8%	4.08	11
Standard GA	0.21	1.50%	10.86	146
GA-heuristic	0.10	2.20%	26.50	53
GA-heuristic with local search	0.08	2.20%	26.50	70
EbFSA_v1	0.06	1.46%	12.04	57
EbFSA_v2	0.07	1.58%	15.28	131

In Table 6.4, PLS means partial least squares which is the most classical technique for variable selection (Section 2.3.1). SPA-MLR is the successive projections algorithm (Section 2.3.2). BVS is the Bayesian variable selection method, and FA-MLR is the firefly algorithm adapted for variable selection. It is important to highlight that the implementations of all such algorithms are published by the author in Paula *et al.* [85].

The outcomes in Table 6.4 reassure the use of the heuristic strategy in GA (Algorithm 4.2) becomes important to reduce the model prediction error in terms of RMSEP. For example, the RMSEP value obtained by GA-heuristic was considerably smaller than standard GA. In addition, although increasing the number of selected variables, the use of GA-heuristic with local search (Algorithm 4.4) provided an even more reduced RMSEP value. Nevertheless, EbFSA_v1 selected the most informative variables and provided the lowest RMSEP despite selecting a relatively large number of variables [10].

Table 6.5 shows the same comparison using different algorithms applied into dataset 2:

[10] It is noteworthy some scientists may prefer a reduced number of variables. In this case, the use of Firefly Algorithm (FA-MLR [85]) would be a more suitable choice.

Table 6.5: *Outcomes comparison between different algorithms using dataset 2.*

Algorithm	RMSEP	MAPE	PRESS	Num. of variables
Participant 1 (IDRC) [55]	0.0140	Not inf.	0.0071	Not inf.
Participant 2 (IDRC) [55]	0.0170	Not inf.	0.0104	Not inf.
Participant 3 (IDRC) [55]	0.0170	Not inf.	0.0104	Not inf.
Participant 4 (IDRC) [55]	0.0310	Not inf.	0.0345	Not inf.
IntGA-MLR [106]	0.0019	6.65%	0.0001	10
Standard GA	0.0291	3.32%	0.0304	68
GA-heuristic	0.0134	5.69%	0.0064	53
GA-heuristic with local search	0.0103	5.69%	0.0038	52
EbFSA_v1	0.0050	3.78%	0.0009	249
EbFSA_v2	0.0018	6.56%	0.0001	7

In Table 6.5, the participants represent four different methods used in the software competition of the International Diffuse Reflectance Conference (IDRC) 2014 [55]. Information such as MAPE, PRESS and number of selected variables were not informed by the conference [11]. IntGA-MLR is a GA implementation based on integer representation. It is important to highlight that the IntGA-MLR is published by the author in Sousa *et al.* [106].

With the overall best statistics Participant 3 won the software competition, but our GA-heuristic and GA-heuristic with local search provided better outcomes. Moreover, IntGA-MLR was able to overcome our GA (with heuristic and local search) and EbFSA_v1. However, EbFSA_v2 was superior to all other algorithms yielding the smallest number of selected variables and the lowest RMSEP value. Such results demonstrate our EbFSA_v2 selected the most informative variables from dataset 2 without decomposing the problem and avoiding schemata disruption. Consequently, it provided a model with the highest predictive ability.

It is important to emphasize it is beyond the scope of this work to compare differences of statistical significance and computational cost between the algorithms. Finally, the computational times of referenced algorithms from Table 6.4 can be obtained in Paula *et al.* [85].

[11] Then, we calculated PRESS information from the RMSEP values.

6.6 Summary

This chapter presented and discussed all experimental results. Section 6.1 argued about the non-decomposability assumption raised up and claimed by Hypothesis 1 (Section 4.1.2). The obtained outcomes demonstrate considerable evidences that in fact the variable selection procedure in multivariate calibration is usually a high dimensionality problem and can not be properly decomposed due to multicollinearity. Figures 6.1 and 6.5 highlight the high level of correlation among the variables from dataset 1 and dataset 2, respectively. It is possible to notice in the charts a brighter region over the search space, which certifies the constant presence of multicollinearity in the datasets.

Section 6.2 presented a schemata disruption analysis by using standard GA implementation with and without the proposed heuristics. When using crossover operator, it becomes clear the constant schemata disruption caused by such recombination operator. Figure 6.8 highlights the trend in raising children worse than parents, while the number of better individuals decrease over the generations. When the heuristics are applied, the results are relatively improved. While one heuristic generates the best initial solutions, the other one tries to identify possible schemata in the population. On the other hand, the proposed local search-based operator is able to provide the best outcomes compared to the heuristics approach. This is due to the fact that such local search operator is simple and does not use recombination. Consequently, the schemata disruption tends to reduce.

Finally, Sections 6.3 and 6.4 described the main results obtained by the proposed Epistasis-based FSA (versions 1 and 2, respectively). These algorithms do not use a population of individuals. On the contrary, such deterministic approaches use only one chromosome and initially select the most linearly independent variables based on the Pearson's linear correlation coefficients matrix, which is assumed as the epistatic relation among variables. EbFSA does not decompose the problem by keeping correlated variables together in the same subset. As a consequence, the schemata disruption is avoided. Thus, EbFSA is able to select the most informative variables and provide the best outcomes when compared with traditional methods from literature.

CHAPTER 7

Conclusions

Based on concepts of decomposability, a problem is said to be decomposable if it is possible to break it into smaller subproblems. For this to happen, all variables making up the problem should not interact with each other and they should be separately treated. However, often in multivariate calibration there are considerable linear dependency among decision variables from spectral data, and this issue indicates such problem can not be properly decomposed. Thus, based on concepts of decomposability, this work aimed to claim that selecting variables in multivariate calibration can be considered as a non-completely decomposable problem due to the constant presence of multicollinearity among variables, which states our first hypothesis.

It is known that recombination operators are able to cause building blocks disruption in GAs by splitting individuals chromosome to form offspring. Current literature lacks a property analysis which consistently clarifies the BBs disruption in GAs. Disrupted BBs are not preserved in new individuals and this leads the algorithm to yield poor performance. In this context, based on schema theory, this work also aimed to claim that not necessarily there exists building blocks formation in spectral data from multivariate calibration due to the high data dimensionality. Even so, schemata disruption caused by crossover operator in GAs affects the non-decomposability assumption of variable selection in multivariate calibration, which assumes our second hypothesis. It is important to highlight that our deep research and experimental results provided significant evidences about the both hypotheses feasibility.

Additionally, we proposed a GA implementation with two heuristics and one simple local search operator. Such heuristics aimed to explore the research empiricism and yielded viable results, but the local search operator was superior. Moreover, two different versions of a novel approach for variable selection were proposed. It was called Epistasis-based Feature Selection Algorithm (EbFSA). EbFSA was able to avoid schemata disruption, select the most informative variables, provide the best outcomes and overcome some state-of-the-art algorithms.

7.1 Summary of Contributions

This doctorate thesis can be summarized based on its contributions such as:

- Proposal of Hypothesis 1:
  - Comprehensive bibliographical review about concepts of decomposability;
  - Development of Equation (4-2);
  - Presentation of three numerical examples;
- Proposal of Hypothesis 2:
  - Broad and deep research about schema theory;
  - Use of Proposition 1 (Section 4.2.2);
  - Presentation of three numerical examples;
- Standard GA implementation with two proposed heuristics:
  - Heuristic for initial solutions generation;
  - Heuristic for possible schemata identification;
- Proposal of a simple local search operator in the GA implementation:
  - Strategy based on the Variable Neighborhood Search method [94];
  - Better outcomes than standard GA and heuristics;
- Proposal of two different versions of a novel approach for variable selection in multivariate calibration called Epistasis-based FSA:
  - Better than local search operator;
  - Superior to some traditional algorithms in terms of number of variables and model predictive ability.
- Published and presented results (including submitted manuscripts):
  - Results of Hypothesis 1 published and presented in the Genetic and Evolutionary Computation Conference (GECCO) 2017 [83];
  - Results of Hypothesis 2 published and presented in the GECCO 2016 [84];
  - Results of an earlier version of this PhD thesis published in the Doctoral Symposium of the EPIA Conference on Artificial Intelligence 2017 [76];
  - Results of Epistasis-based FSA submitted to a relevant scientific journal.

All details about the published and elaborated works as well as those accepted for publication during the course of this thesis are listed below. The lists of items are listed in ascending order of year of publication.

7.1.1 Published papers

1. 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 (**Qualis A1**), Denver - GECCO '16 Companion. New York: ACM Press, 2016. p. 1461-1462 [78].
2. 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 (**Qualis A1**), Denver - GECCO '16 Companion. New York: ACM Press. p. 1031-1034 [84].
3. Variable Selection as a Non-Completely Decomposable Problem: A Case Study in Multivariate Calibration. In: 2017 on Genetic and Evolutionary Computation Conference (**Qualis A1**), Berlin - GECCO '17 Companion. New York: ACM Press. p. 1399-1402 [83].
4. Variable Selection in Multivariate Calibration considering Non-Decomposability Assumption and Building Blocks Hypothesis. In: Doctoral Symposium of the 18th EPIA Conference on Artificial Intelligence (**Qualis B2**), Porto - Portugal. Proceedings of EPIA 2017. p. 44-55 [76].

7.1.2 Main published related papers

1. A GPU-Based Implementation of the Firefly Algorithm for Variable Selection in Multivariate Calibration Problems. Plos One (**Qualis A1**), v. 9, p. e114145, 2014 [85].
2. Multiobjective Firefly Algorithm for Variable Selection in Multivariate Calibration. In: Lecture Notes in Computer Science (**Qualis B2**), v. 9273, p. 274-279, 2015 [82].
3. A Metaheuristic Implementation for Variable Selection in Multivariate Calibration Models. In: 11th Metaheuristics International Conference (**Qualis B3**), Agadir. Proceedings of MIC 2015 [81].
4. Parallel regressions for variable selection using GPU. Computing - Springer (**Qualis B1**), v. 98, p. 1-16, 2016 [79].
5. A Compact Firefly Algorithm for the Variable Selection Problem in Pharmaceutical Ingredient Determination. In: 2016 IEEE Congress on Evolutionary Computation (**Qualis A1**), Vancouver. Proceedings of IEEE CEC, 2016. p. 3832-3838 [77].
6. Modern Metaheuristic with Multi-Objective Formulation for the Variable Selection Problem. Journal of Computer Science (**Qualis B1**), v. 13, p. 659-666, 2017 [80].
7. Integer-based Genetic Algorithm for Feature Selection in Multivariate Calibration. In: 2017 IEEE Congress on Evolutionary Computation (**Qualis A1**), San Sebastian, Spain. Proceedings of IEEE CEC, 2017. p. 2315-2320 [106].

7.1.3 Manuscripts in peer review process

1. Epistasis-based FSA: Two Versions of A Novel Approach for Variable Selection in Multivariate Calibration. Submitted in: Engineering Applications of Artificial Intelligence (**Qualis A2**).

7.1.4 Manuscripts in written process

1. Genetic Algorithm: A Poor Technique for Variable Selection in Multivariate Calibration. To be submitted in: Analytica Chimica Acta (**Qualis A1**).

7.1.5 Main awards

1. 2014 - Second place in the Contests of Thesis and Dissertations in Computer Architecture and High Performance Computing (WSCAD-CTD 2014), promoted by the Brazilian Society of Computing [75].
2. 2015 - Recognition certificate from University Council (CONSUNI) of the Federal University of Goiás, Brazil.
3. 2017 - Grant of USD 700.00 from the Association for Computing Machinery (ACM) for participation and paper presentation at the Genetic and Evolutionary Computation Conference (GECCO) 2017, in Berlin [83].

7.1.6 Countries where author presented scientific papers

1. Morocco - Metaheuristic International Conference 2015 [81].
2. USA - Genetic and Evolutionary Computation Conference 2016 [84].
3. Canada - IEEE Congress on Evolutionary Computation 2016 [77].
4. Germany - Genetic and Evolutionary Computation Conference 2017 [83].

7.2 Limitations of Our Proposal

We consider our both hypotheses and all algorithms proposed in this work are useful. However, they suffer from some limitations when considering more general applications. Such limitations are now discussed.

In the context of this work, it is considered only the binary-coded Genetic Algorithm. This unique representation narrows the applicability of both hypotheses. Thus, Hypothesis 1 and Hypothesis 2 can not be generalized. They lack some theorem to be general justified. A possible solution for this issue could be the proposal of one theorem together with a mathematical proof in order to demonstrate they can be applied to other optimization problems with high data dimensionality as well as to other types of representation.

Another limitation concerns the proposed local search operator. Although our local search operator is able to provide viable outcomes, it was implemented on a simple manner. It is based on the Variable Neighborhood Search method [94]. We believe it can be further improved by applying a more efficient strategy. Consequently, better results may be achieved.

The Epistasis-based Feature Selection Algorithm (EbFSA) uses the concept of epistasis considering the Pearson's linear correlation coefficient as the epistatic relation among the variables in the dataset. Nevertheless, it is not improved by mixing concepts from other theories. For example, we consider the use of Linkage Learning (Section 3.5.2) may provide additional information about genes (variables) interdependence. Moreover, the use of Mutation-based Compact Genetic Algorithm [101] may perform a more significant exploitation and exploration in the search space, possibly yielding even better outcomes.

EbFSA_v1 was superior to EbFSA_v2 regarding dataset 1, and EbFSA_v2 overcomed EbFSA_v1 regarding dataset 2. We presume the strategy used by EbFSA_v2 tends to be better when applied in even larger datasets with high data dimensionality. This may happen because EbFSA_v2 selects and maintains in the calibration model the most informative variables on an individually manner. On the other hand, the second stage of EbFSA_v1 discards two correlated variables previously selected if the prediction error is not reduced. Then, one of the two variables may be relevant together with other variables but it is discarded anyway. However, the use of additional datasets and a deeper investigation become necessary to demonstrate such assumption.

Finally, we suggest future work in the next section which could be carried out to address these limitations and discuss some potential improvements to our proposed hypotheses and algorithms.

7.3 Future Work

For future work, we aim to propose at least one theorem to formally demonstrate the variable selection procedure in multivariate calibration is a non-decomposable problem (Hypothesis 1) as well as the schemata disruption caused by crossover operators affects the non-decomposability assumption (Hypothesis 2). The use of some theorem with mathematical proof could be an important manner to prove that indeed recombination operators are an infeasible technique for variable selection in multivariate calibration models as well as for other optimization problems with high data dimensionality.

We have an additional hypothesis (maybe called Hypothesis 3) which assumes that spectral regions reliably representing the property of interest in the analyzed sample can be set in blocks of variables. Section 3.4.3 tried to point out this possible assumption. Such blocks could be considered as independent blocks or correlated blocks. In this case, we aim to use some additional linear correlation measure to assess spectral orthogonality among variables in order to demonstrate the possible existence of such blocks in the spectral regions. Another possible alternative could be the use of concepts and tools from analytical chemistry to measure chemical interdependence between two or more wavelengths absorbed by molecules in the sample.

We also aim to propose an enhanced Epistasis-based FSA (EbFSA) in order to reduce the number of selected variables without incurring loss of performance in terms of model prediction error. In this sense, multi-objective optimization problems may be tackled accordingly. Moreover, EbFSA may be adapted and applied to solve classification problems. Therefore, the two versions of EbFSA can be assessed in order to verify which one is the best in such context, and a suitable comparison between classical algorithms could be performed.

Finally, Genetic Algorithms (GAs) coded with other types of representation may be investigated inside the context of the second hypothesis. For instance, the integer-based GA (IntGA-MLR [106]) may be analyzed in order to verify if there is the possibility of schemata (or building blocks) formation (even without binary codification) as well as if schemata can be disrupted in such representation.

References

[1] AHN, C. W.; RAMAKRISHNA, R. S. **Elitism-based compact genetic algorithms**. *Evolutionary Computation, IEEE Transactions on*, 7:367–385, 2003.

[2] AHN, C. W. **Advances in evolutionary algorithms**. Springer, 2006.

[3] ALLEN, M. P. **The Problem of Multicollinearity In: Understanding Regression Analysis**. Springer Science and Business Media, 2004.

[4] ALTENBERG, L. **The schema theorem and prices theorem**. *Foundations of genetic algorithms*, 3:23–49, 1995.

[5] ARAKAWA, M.; YAMASHITA, Y.; FUNATSU, K. **Genetic algorithm-based wavelength selection method for spectral calibration**. *Journal of Chemometrics*, 25(1):10–19, 2011.

[6] ARAÚJO, M. C. U.; SALDANHA, T. C. B.; GALVÃO, R. K. H.; YONEYAMA, T.; CHAME, H. C.; VISANI, V. **The successive projections algorithm for var. selec. in spect. multicomp. anal.** *Chemometrics and Intelligent Laboratory Systems*, 57(2):65–73, 2001.

[7] ATKINS, P.; PAULA, J. **Atkins Physical Chemistry**. Oxford University Press, 2002.

[8] ATKINS, P.; PAULA, J. **Elements of physical chemistry**. Oxford University Press, 2013.

[9] BEEBE, K. R.; PELL, R. J.; SEASHOLTZ, M. B. **Chemometrics: a practical guide**. 1998.

[10] BOURBAKI, N. **Algebra II**. Springer, 2003.

[11] BRADSTREET, R. B. **The kjeldahl method for organic nitrogen.** *The Kjeldahl method for organic nitrogen.*, 1965.

[12] BREITKREITZ, M. C.; RAIMUNDO, I. M.; ROHWEDDER, J. J. R.; PASQUINI, C.; FILHO, H. A. D.; JOSE, G. E.; ARAUJO, M. C. U. **Determination of total sulfur**

in diesel fuel employing nir spectroscopy and multivariate calibration. *The Analyst*, 128:1204–1207, 2003.

[13] BRIDGES, C. L.; GOLDBERG, D. E. **An analysis of reproduction and crossover in a binary-coded genetic algorithm**. *Grefenstette*, 878:9–13, 1987.

[14] BROWN, S. D.; BLANK, T. B.; SUM, S. T.; WEYER, L. G. **Chemometrics**. *Analytical chemistry*, 66(12):315–359, 1994.

[15] CHAN, K. Y.; AYDIN, M. E.; FOGARTY, T. C. **An epistasis measure based on the analysis of variance for the real-coded representation in genetic algorithms**. In: *Evolutionary Computation, 2003. CEC'03. The 2003 Congress on*, volume 1, p. 297–304. IEEE, 2003.

[16] CHEN, Y.-P. **Linkage learning genetic algorithm**. In: *Extending the Scalability of Linkage Learning Genetic Algorithms*, p. 35–43. Springer, 2006.

[17] CHEN, Y.-P.; GOLDBERG, D. E. **Convergence time for the linkage learning genetic algorithm**. *Evolutionary computation*, 13(3):279–302, 2005.

[18] CHONG, I.-G.; JUN, C.-H. **Performance of some variable selection methods when multicollinearity is present**. *Chemometrics and Intelligent Laboratory Systems*, 78(1):103–112, 2005.

[19] CHUNG, W. S.; PEREZ, R. A. **The schema theorem considered insufficient**. In: *Proc. of Sixth Int. Conf. on Tools with Art. Intell.*, p. 748–751. 1994.

[20] CIBAS, T.; SOULIÉ, F. F.; GALLINARI, P.; RAUDYS, S. **Variable selection with neural networks**. *Neurocomputing*, 12(2):223–248, 1996.

[21] CLEVELAND, W. S.; DEVLIN, S. J. **Locally weighted regression: an approach to regression analysis by local fitting**. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

[22] COELLO, C. A. C.; VAN VELDHUIZEN, D. A.; LAMONT, G. B. **Evolutionary algorithms for solving multi-objective problems**, volume 242. Springer, 2002.

[23] CORTINA, J. M. **Interaction, nonlinearity, and multicollinearity: Implications for multiple regression**. *Journal of Management*, 19(4):915–922, 1993.

[24] DAS, S. **Filters, wrappers and a boosting-based hybrid for feature selection**. In: *ICML*, volume 1, p. 74–81. Citeseer, 2001.

[25] DASH, M.; LIU, H. **Feature selection for classification**. *Intelligent data analysis*, 1(3):131–156, 1997.

[26] DAVIDOR, Y. **Epistasis variance: Suitability of a representation to genetic algorithms**. *Complex Systems*, 4(4):369–383, 1990.

[27] DE REFLECTÂNCIA DIFUSA, C. I. http://www.idrc-chambersburg.org/shootout.html, 2008.

[28] DROSTE, S. **A rigorous analysis of the compact genetic algorithm for linear functions**. *Natural Computing*, 5:257–283, 2006.

[29] ESHELMAN, L. J. **The chc adaptive search algorithm: How to have safe search when engaging**. *Foundations of Genetic Algorithms 1991 (FOGA 1)*, 1:265, 2014.

[30] FAN, K.-C.; YU, T.-L.; LEE, J.-T. **Interaction detection by nfe estimation: a practical view of building blocks**. In: *Proceedings of the 13th annual conference companion on Genetic and evolutionary computation (2011)*, p. 71–72, 2011.

[31] FARRAR, D. E.; GLAUBER, R. R. **Multicollinearity in regression analysis: The problem revisited**. *The Review of Economics and Statistics*, 49(1):92–107, 1967.

[32] FERRAND, M.; HUQUET, B.; BARBEY, S.; BARILLET, F.; FAUCON, F.; LARROQUE, H.; LERAY, O.; TROMMENSCHLAGER, J.; BROCHARD, M. **Determination of fatty acid profile in cow's milk using mid-infrared spectrometry: Interest of applying a variable selection by genetic algorithms before a pls regression**. *Chemometrics and Intelligent Laboratory Systems*, 106(2):183–189, 2011.

[33] FERREIRA, M. M. C.; ANTUNES, A. M.; MELGO, M. S.; VOLPE, P. L. O. **Quimiometria I: calibracao multivariada, um tutorial**. *Quimica Nova*, 22:724–731, 09 1999.

[34] FILHO, A. R. G.; GALVÃO, R. K. H.; ARAÚJO, M. C. U. **Effect of the subsampling ratio in the application of subagging for multivariate calibration with the successive projections algorithm**. *Journal of the Brazilian Chemical Society*, 22:2225–2233, 11 2011.

[35] FOGEL, D. B.; GHOZEIL, A. **Schema processing under proportional selection in the presence of random effects**. *Evolutionary Computation, IEEE Transactions on*, 1(4):290–293, 1997.

[36] FONLUPT, C.; ROBILLIARD, D.; PREUX, P. **A bit-wise epistasis measure for binary search spaces**. In: *Parallel Problem Solving from Natureâ"PPSN V*, p. 47–56. Springer, 1998.

[37] GALVÃO FILHO, A. R.; GALVÃO, R. K.; ARAÚJO, M. C. U. **Effect of the subsampling ratio in the application of subagging for multivariate calibration with**

the successive projections algorithm. *Journal of the Brazilian Chemical Society*, 22(11):2225–2233, 2011.

[38] GELADI, P.; MARTENS, H. **A calibration tutorial for spectral data. part 1. data pretreatment and principal component regression using matlab**. *Journal of Near Infrared Spectroscopy*, 4:225, 1996.

[39] GEORGE, E. I. **The variable selection problem**. *Journal of the American Statistical Association*, 95(452):1304–1308, 2000.

[40] GOLDBERG, D. E. **Genetic algorithms and walsh functions: Part i, a gentle introduction**. *Complex systems*, 3(2):129–152, 1989.

[41] GOLDBERG, D. E. **The design of innovation: Lessons from and for competent genetic algorithms**, volume 7. Springer Science and Business Media, 2002.

[42] GOLDBERG, D. E.; DEB, K. **A comparative analysis of selection schemes used in genetic algorithms**. *Foundations of genetic algorithms*, 1:69–93, 1991.

[43] GOLDBERG, D. E.; DEB, K.; CLARK, J. H. **Genetic algorithms, noise, and the sizing of populations**. *Complex systems*, 6:333–362, 1991.

[44] GOLDBERG, D. E.; DEB, K.; THIERENS, D. **Toward a better understanding of mixing in genetic algorithms.** *J. Soc. Instrument and Control Engineers*, 32(1):10–16, 1993.

[45] GOLDBERG, D. E.; SASTRY, K. **A practical schema theorem for genetic algorithm design and tuning**. In: *Proceedings of the genetic and evolutionary computation conference (2001)*, p. 328–335, 2001.

[46] GOLDBERG, D. E.; SASTRY, K.; LATOZA, T. **On the supply of building blocks**. In: *Proceedings of the Genetic and Evolutionary Computation Conference (2001)*, p. 336–342, 2001.

[47] GOLDBERG, D. E. **A note on the disruption due to crossover in a binary-coded genetic algorithm**. University of Alabama, 1987.

[48] GUYON, I.; ELISSEEFF, A. **An introduction to variable and feature selection**. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[49] HAALAND, D. M.; THOMAS, E. V. **Partial least-squares methods for spectral analyses. 1. relation to other quantitative calibration methods and the extraction of qualitative information**. *Analytical Chemistry*, 60(11):1193–1202, 1988.

[50] HARIK, G. **Linkage learning via probabilistic modeling in the ecga**. *Urbana*, 51(61):801–817, 1999.

[51] HARIK, G. R.; GOLDBERG, D. E. **Learning linkage.** In: *FOGA*, volume 4, p. 247–262, 1996.

[52] HARIK, G. R.; LOBO, F. G.; GOLDBERG, D. E. **The compact genetic algorithm**. *Evolutionary Computation, IEEE Transactions on*, 3(4):287–297, 1999.

[53] HOLLAND, J. H. **Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.** 1975.

[54] HOQUE, M. T.; CHETTY, M.; DOOLEY, L. S. **Generalized schemata theorem incorporating twin removal for protein structure prediction**. In: *IAPR International Workshop on Pattern Recognition in Bioinformatics*, p. 84–97. Springer, 2007.

[55] IGNE, B.; BOGOMOLOV, A.; BU, D.; DARDENNE, P.; GALYANIN, V.; TILLMANN, P. **Summary of the 2014 idrc software shoot-out**. *NIR news*, 26(2):8–14, 2015.

[56] KALIVAS, J. **Mathematical analysis of spectral orthogonality**, volume 17. CRC Press, 1993.

[57] KENNARD, R. W.; STONE, L. A. **Computer aided design of experiments**. *Technometrics*, 11(1):137–148, 1969.

[58] KENT, J. T. **Information gain and a general measure of correlation**. *Biometrika*, 70(1):163–173, 1983.

[59] KOHAVI, R.; JOHN, G. H. **Wrappers for feature subset selection**. *Artificial intelligence*, 97(1):273–324, 1997.

[60] LAWSON, C. L.; HANSON, R. J. **Solving least squares problems**, volume 161. SIAM, 1974.

[61] LEARDI, R. **Genetic algorithms in chemometrics and chemistry: a review**. *Journal of Chemometrics*, 15(7):559–569, 2001.

[62] LEE, H.; YU, T.-L. **Off-line building block identification: detecting building blocks directly from fitness without genetic algorithms**. In: *Proceedings of the 14th annual conference on Genetic and evolutionary computation (2012)*, p. 641–648, 2012.

[63] LEE RODGERS, J.; NICEWANDER, W. A. **Thirteen ways to look at the correlation coefficient**. *The American Statistician*, 42(1):59–66, 1988.

[64] LUCENA, D. V.; LIMA, T. W.; SOARES, A. S.; DELBEM, A. C. B.; GALVAO, A. R.; COELHO, C. J.; LAUREANO, G. T. **Multi-objective evolutionary algorithm for variable selection in calibration problems: A case study for protein concentration prediction**. In: *Proceedings of 2013 IEEE Congress on Evolutionary Computation*, p. 1053–1059, 2013.

[65] MAN, K.-F.; TANG, K. S.; KWONG, S. **Genetic algorithms: Concepts and designs**. Springer Science and Business Media, 2012.

[66] MARTENS, H. **Multivariate calibration**. John Wiley & Sons, 1991.

[67] MASON, A. **A non-linearity measure of a problem's crossover suitability**. In: *Evolutionary Computation, 1995, IEEE International Conference on*, volume 1, p. 68. IEEE, 1995.

[68] MIQUELINA, P. F. P. **Visualization of genetic algorithm operation on additive decomposable functions**. Master's thesis, University of Algarve, Portugal, 2013.

[69] MITCHELL, M.; HOLLAND, J.; FORREST, S. **Relative building-block fitness and the building block hypothesis**. *D. Whitley, Foundations of Genetic Algorithms*, 2:109–126, 1993.

[70] MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. John Wiley & Sons, 2015.

[71] MURPHY, A. H. **Skill scores based on the mean square error and their relationships to the correlation coefficient**. *Monthly weather review*, 116(12):2417–2424, 1988.

[72] NAES, T.; MEVIK, B.-H. **Understanding the collinearity problem in regression and discriminant analysis**. *Journal of Chemometrics*, 15(4):413–426, 2001.

[73] NEWMAN, D. R. **The use of linkage learning in genetic algorithms**, 2006.

[74] NIAZI, A.; LEARDI, R. **Genetic algorithms in chemometrics**. *Journal of Chemometrics*, 26(6):345–351, 2012.

[75] PAULA, L. C. M. **Paralelizacao de algoritmos aps e firefly para selecao de variaveis em problemas de calibracao multivariada**. Master's thesis, Universidade Federal de Goias, 2014. Programa de Pos-graduacao em Ciencia da Computacao (UFG).

[76] PAULA, L. C. M. **Variable selection in multivariate calibration considering non-decomposability assumption and building blocks hypothesis**. In: *Doctoral*

Symposium of the 18th EPIA Conference on Artificial Intelligence, p. 44–55. EPIA, 2017.

[77] PAULA, L. C. M.; NOGUEIRA, H. V.; SOARES, A. S.; LIMA, T. W.; COELHO, C. J. **A compact firefly algorithm for the variable selection problem in pharmaceutical ingredient determination**. In: *Proceedings of IEEE Congress on Evolutionary Computation*, 2016.

[78] PAULA, L. C. M.; SOARES, A.; LIMA, T.; COELHO, C. **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*, p. 1461–1462. ACM, 2016.

[79] PAULA, L. C. M.; SOARES, A.; SOARES, T.; FILHO, A.; COELHO, C.; DELBEM, A.; MARTINS, W. **Parallel regressions for variable selection using gpu**. *Computing*, p. 1–16, 2016.

[80] PAULA, L. C. M.; SOARES, A.; SOARES, T.; OLIVEIRA, A.; COELHO, C. **Modern metaheuristic with multi-objective formulation for the variable selection problem**. *Journal of Computer Science*, 13:659–666, 2017.

[81] PAULA, L. C. M.; SOARES, A. S. **A metaheuristic implementation for variable selection in multivariate calibration models**. In: *Proceedings of the Metaheuristic International Conference*.

[82] PAULA, L. C. M.; SOARES, A. S. **Multiobjective firefly algorithm for variable selection in multivariate calibration**. In: *Progress in Artificial Intelligence*, p. 274–279. Springer, 2015.

[83] PAULA, L. C. M.; SOARES, A. S.; LIMA, T. W.; COELHO, C. J. **Variable selection as a non-completely decomposable problem: A case study in multivariate calibration**. In: *Proceedings of the 2017 on Genetic and Evolutionary Computation Conference Companion*, p. 1399–1402. ACM, 2017.

[84] PAULA, L. C. M.; SOARES, A. S.; LIMA, T. W.; COELHO, C. J.; FILHO, A. R. G. **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 Companion*, p. 1031–1034. ACM, 2016.

[85] PAULA, L. C. M.; SOARES, A. S.; LIMA, T. W.; DELBEM, A. C. B.; COELHO, C. J.; FILHO, A. R. G. **A gpu-based implementation of the firefly algorithm for vari-**

able selection in multivariate calibration problems. *Plos One*, 9(12):e114145, 2014.

[86] PAULA, L. C. M.; SOARES, A. S.; LIMA, T. W.; DELBEM, A. C. B.; COELHO, C. J.; FILHO, A. R. G. **Parallelization of a modified firefly algorithm using gpu for variable selection in a multivariate calibration problem**. *International Journal of Natural Computing Research*, 4:31–42, 2014.

[87] PAULA, L. C. M.; SOARES, A. S.; LIMA, T. W.; MARTINS, W. S.; FILHO, A. R. G.; COELHO, C. J. **Partial parallelization of the successive projections algorithm using compute unified device architecture**. In: *International Conference on Parallel and Distributed Processing Techniques and Applications*, p. 737–741. 2013.

[88] PAWAR, R.; SAINI, J. **Schemata tutorial**. *Journal of Multi Disciplinary Engineering Technologies*, 9(2):25–36, 2015.

[89] PELIKAN, M.; GOLDBERG, D. E.; CANTU-PAZ, E. **Linkage problem, distribution estimation, and bayesian networks**. *Evolutionary computation*, 8(3):311–340, 2000.

[90] PLACKETT, R. L. **Karl pearson and the chi-squared test**. *International Statistical Review/Revue Internationale de Statistique*, p. 59–72, 1983.

[91] POLI, R. **Exact schema theory for gen. progr. and var.-length gen. alg. with one-point crossover**. *Genetic Progr. and Evolvable Mach.*, 2(2):123–163, 2001.

[92] RADCLIFFE, N. J. **Forma analysis and random respectful recombination.** In: *ICGA*, volume 91, p. 222–229, 1991.

[93] ROTHLAUF, F. **Representations for genetic and evolutionary algorithms**. Springer-Verlag, Berling, Heidelberg, Netherlands, 2006.

[94] ROTHLAUF, F. **Design of modern heuristics: principles and application**. Springer Science & Business Media, 2011.

[95] ROY, P. P.; ROY, K. **On some aspects of variable selection for partial least squares regression models**. *QSAR & Combinatorial Science*, 27(3):302–313, 2008.

[96] RUIZ-LEÓN, J. **Decoupling with stability of linear multivariable systems: An algebraic approach**. *Latin American applied research*, 34:179–186, 2004.

[97] RUMMEL, R. J. **Understanding correlation**. *Honolulu: Department of Political Science, University of Hawaii*, 1976.

[98] SASTRY, K.; GOLDBERG, D. E. **On extended compact genetic algorithm**. In: *Late-Breaking Paper at the Genetic and Evolutionary Computation Conference*, p. 352–359. 2000.

[99] SASTRY, K.; GOLDBERG, D. E. **Probabilistic model building and competent genetic programming**. In: *Genetic Programming Theory and Practice*, p. 205–220. Springer, 2003.

[100] SKOOG, D. A.; HOLLER, F. J.; NIEMAN, T. A. **Principles of instrumental analysis**. 1998.

[101] SOARES, A. S.; LIMA, T. W. **Mutation-based compact ga for spectroscopy variable selection in determining protein concentration in wheat grain**. *Elec. Letters*, 50:932–934, 2014.

[102] SOARES, A.; LIMA, T. W.; LUCENA, D. V.; SALVINI, R. L.; LAUREANO, G. T.; COELHO, C. J. **Spectroscopic multicomponent analysis using multi-objective optimization for variable selection**. *Computer Technology and Application*, 4(9), 2013.

[103] SOARES, A. S.; FILHO, A. R. G.; GALVÃO, R. K. H.; ARAÚJO, M. C. U. **Improving the computational efficiency of the successive projections algorithm by using a sequential regression implementation: a case study involving nir spectrometric analysis of wheat samples**. *Journal of the Brazilian Chemical Society*, 21(4):760–763, 2010.

[104] SOARES, A. S.; GALVÃO, R. K. H.; ARAÚJO, M. C. U.; SOARES, S. F. C.; PINTO, L. A. **Multi-core computation in chemometrics: case studies of voltammetric and nir spectrometric analyses**. *Journal of the Brazilian Chemical Society*, 21:1626–1634, 2010.

[105] SOARES, S. F. C.; GOMES, A. A.; ARAUJO, M. C. U.; GALVÃO, R. K. H.; OTHERS. **The successive projections algorithm**. *TrAC Trends in Analytical Chemistry*, 42:84–98, 2013.

[106] SOUSA, R. S.; DE LIMA, T. W.; PAULA, L. C. M.; LIMA, R. L.; ARLINDO FILHO, R.; SOARES, A. S. **Integer-based genetic algorithm for feature selection in multivariate calibration**. In: *Evolutionary Computation (CEC), 2017 IEEE Congress on*, p. 2315–2320. IEEE, 2017.

[107] STEPHENS, C.; WAELBROECK, H. **Schemata evolution and building blocks**. *Evolutionary computation*, 7(2):109–124, 1999.

[108] STEPHENS, C.; WAELBROECK, H.; AGUIRRE, R. **Schemata as building blocks: Does size matter**. *Foundations of Genetic Algorithms*, 5:117–133, 1999.

[109] TALAVERA, L. **An evaluation of filter and wrapper methods for feature selection in categorical clustering**. In: *International Symposium on Intelligent Data Analysis*, p. 440–451. Springer, 2005.

[110] TARPEY, T. **A note on the prediction sum of squares statistic for restricted least squares**. *The American Statistician*, 54(2):116–118, 2000.

[111] TOBIAS, R. **An introduction to partial least squares regression**. In: *Proceedings of Ann. SAS Users Group Int. Conf., 20th, Orlando, FL*, p. 2–5, 1995.

[112] UNCU, Ö.; TÜRKŞEN, I. **A novel feature selection approach: combining feature wrappers and filters**. *Information Sciences*, 177(2):449–466, 2007.

[113] VAN DER MEER, F. **The effectiveness of spectral similarity measures for the analysis of hyperspectral imagery**. *International journal of applied earth observation and geoinformation*, 8(1):3–17, 2006.

[114] WALCZAK, B.; MASSART, D. **The radial basis functions partial least squares approach as a flexible non-linear regression technique**. *Analytica Chimica Acta*, 331(3):177–185, 1996.

[115] WATSON, R. A.; HORNBY, G. S.; POLLACK, J. B. **Modeling building-block interdependency**. In: *International Conference on Parallel Problem Solving from Nature*, p. 97–106. Springer, 1998.

[116] WESTAD, F.; MARTENS, H. **Variable selection in near infrared spectroscopy based on significance testing in partial least squares regression**. *Journal of Near Infrared Spectrosc.*, 8:117–124, 2000.

[117] WHITLEY, L. D. **Foundations of genetic algorithms 2**, volume 2. Morgan Kaufmann, 1993.

[118] WRIGHT, A. H. **The exact schema theorem**. *arXiv preprint arXiv:1105.3538*, 2011.

[119] XIAOBO, Z.; JIEWEN, Z.; POVEY, M. J.; HOLMES, M.; HANPIN, M. **Variables selection methods in near-infrared spectroscopy**. *Analytica chimica acta*, 667(1):14–32, 2010.

[120] XING, H.; QU, R. **A compact genetic algorithm for the network coding based resource minimization problem**. *Applied Intelligence*, 36:809–823, 2012.

[121] XU, H.; QI, B. **Variable selec. in vis. and nir spectra: Appl. to on-line determ. of sugar content in pears**. *Journal of Food Engin.*, 109(1):142–147, 2012.

[122] YANG, X. S. **Nature-inspired metaheuristic algorithms**. Luniver Press, 2008.

[123] YANG, X. S. **Firefly algorithm, stochastic test functions and design optimisation**. *International Journal of Bio-Inspired Computation*, 2(2):78–84, 2010.

[124] YUN, Y.-H.; CAO, D.-S.; TAN, M.-L.; YAN, J.; REN, D.-B.; XU, Q.-S.; YU, L.; LIANG, Y.-Z. **A simple idea on applying large regression coefficient to improve the genetic algorithm-pls for variable selection in multivariate calibration**. *Chemometrics and Intelligent Laboratory Systems*, 130:76–83, 2014.