UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

TESE DE DOUTORADO

REDUÇÃO DE DIMENSIONALIDADE PARA DADOS ESPECTRAIS COLINEARES

Felipe Soares

Porto Alegre, Abril de 2022

Felipe Soares

REDUÇÃO DE DIMENSIONALIDADE PARA DADOS ESPECTRAIS COLINEARES

Tese submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do título de Doutor em Engenharia, na área de concentração em Sistemas de Produção.

Orientador: Prof. Michel José Anzanello, *Ph.D.*

.

Porto Alegre,

2022

Felipe Soares

REDUÇÃO DE DIMENSIONALIDADE PARA DADOS ESPECTRAIS COLINEARES

Esta tese foi julgada adequada para a obtenção do título de Doutor em Engenharia e aprovada em sua forma final pelo Orientador e pela Banca Examinadora designada pelo Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul.

_____

Prof. Michel Anzanello, *PhD*
Orientador PPGEP/UFRGS

_____

Prof. Alejandro Germán Frank
Coordenador PPGEP/UFRGS

Banca Examinadora:

Professor Leandro dos Santos Coelho, Dr. (Escola Politécnica – PUCPR/UFPR)

Professor Marcelo Farenzena, Dr. (PPGEQ/UFRGS)

Alessandro Kahmann, Dr. (Campus Litoral Norte/UFRGS)

AGRADECIMENTOS

Agradeço primeiramente a meus pais por todo o esforço, e muitos sacrifícios, que realizaram para que eu tivesse a oportunidade de primeiramente estudar no colégio Tiradentes e então na UFRGS. Também agradeço minha irmã Elenice por sempre ter estado ao meu lado me apoiando (e me aguentando).

Também expresso meus agradecimentos a toda minha família e a todos que de forma direta ou indireta fizeram parte desta tese, em especial:

Às minhas amigas Larissa e Alice, pela amizade de longa data e que sempre estiveram ao meu lado e me ouviram (e ainda ouvem) reclamar inúmeras vezes :P

Ao meu tio Dr. Valdir Araújo, que desde a minha infância foi uma referência em conhecimento e que definitivamente influenciou minha decisão em ser engenheiro e em realizar o doutorado.

Aos meus amigos do PPGEP, especialmente Gabrielli (Japa), Alessandro, Miriam, e Diego. Nossos coffee breaks na salinha do PPGEP, nossas idas ao Xirú, e risadas no LOPP marcaram meu doutorado.

Ao meu orientador, professor Michel Anzanello, pelos ensinamentos, paciência, e parceria durante minha graduação, mestrado e doutorado.

Aos amigos que fiz na Universidade de Sheffield, por nossa caminhada juntos (especialmente durante a pandemia) e nossos momentos de descontração no laboratório.

Ao CNPq, pelo financiamento, o qual foi fundamental para a realização deste doutorado.

It's not an experiment if you know it's going to work.

Jeff Bezos

RESUMO

Na análise de dados, a identificação das variáveis relevantes para uma determinada tarefa de aprendizagem da máquina pode ajudar a construir modelos mais precisos, robustos e explicáveis. Embora avanços recentes em redes neurais, como *autoencoders* e redes neurais profundas, tenham proporcionado abordagens que implicitamente realizam a redução de dimensionalidade, tais modelos usualmente requerem grandes tamanhos de amostra e podem não ser explicáveis, podendo ter aplicabilidade restrita em diversos tipos de bancos de dados, como os de espectroscopia. Bancos de dados espectroscópicos têm como característica um elevado número de variáveis que tendem a ser colineares e geralmente se apoiam em menor número de amostras do que variáveis, o que pode deteriorar o desempenho de diversas técnicas multivariadas aplicadas a tais dados. Desta forma, esta tese propõe métodos de seleção de variáveis aplicados a dados espectroscópicos com o objetivo de realizar agrupamento, classificação e regressão em conjuntos de dados abrangendo diferentes áreas. Esta tese é composta de quatro artigos, três de pesquisa aplicada, e uma comunicação. No primeiro artigo, um índice de importância de variáveis (IIV) é proposto para selecionar os comprimentos de onda mais relevantes para o agrupamento de amostras de acordo com suas similaridades. O IIV proposto é baseado na combinação do escalonamento multidimensional (para redução de dimensionalidade) e análise de Procrustes para derivar uma matriz de projeção. No segundo artigo, com o objetivo de selecionar variáveis para um problema de regressão, outro VII é derivado com base nos pesos da matriz de projeção obtida a partir de uma redução de dimensão através da regressão inversa por fatias localizadas (LSIR). No terceiro artigo, uma comunicação relacionada a um artigo publicado recentemente, foram apontadas falhas de projeto em um experimento com o objetivo de classificar espectros Raman de plasma sanguíneo de pacientes positivos para COVID e controles. Esta comunicação também estabeleceu *baselines* não enviesados para o quarto artigo, no qual o algoritmo de Máxima Relevância Mínima Redundância (mRMR) para seleção de variáveis é melhorado a fim de levar em conta as dependências lineares no conjunto de variáveis selecionadas. O aprimoramento proposto, denominado PCA-mRMR, é aplicado ao mesmo conjunto de dados do terceiro artigo com propósito de classificação. Em todos os três artigos de pesquisa, os métodos propostos foram comparados com abordagens de seleção de variáveis já existentes e seu desempenho foi avaliado.

Palavras-chave: Seleção de variáveis, Classificação, Agrupamento, Regressão, LSIR, PCA, MDS, Espectroscopia.

ABSTRACT

In data analysis, identifying the most relevant features for a given machine learning task can help build more accurate, robust, and explainable models. Although recent advances in neural networks, such as autoencoders and deep neural nets, have provided approaches that implicitly perform dimension reduction, they usually require large sample sizes and may not be explainable. One of such cases is the analysis of spectroscopic data, which is characterised by colinear features (variables or wavelengths) and usually have less samples than features, thus suffering for the curse of dimensionality. Considering this setting, this thesis presents propositions for features election methods applied to spectroscopic data with the goal to perform clustering, classification, and regression in datasets spanning different areas. This thesis is comprised of four articles, three applied research ones, and one communication. In the first article, a feature importance index (FII) is proposed to select the most relevant wavelengths for clustering. This FII is based on the combination of multidimensional scaling (for dimension reduction) and Procrustes analysis to derive a projection matrix. In the second article, with the goal of selecting features for a regression problem, another FII is derived based on the weights of the projection matrix from a Localized Sliced Inverse Regression dimension reduction. In the third article, a communication related to a recent published article, design flaws were pointed out in an experiment aiming to classify Raman spectra of blood plasma of COVID positive patients and controls. This article also established unbiased baselines for the fourth article. In the fourth article, the Maximum Relevancy Minimum Redundancy (mRMR) algorithm for feature selection is improved in order to account for linear dependencies in the selected features. The proposed improved, named PCA-mRMR, is applied to the same dataset of article three, being a classification task. In all three research articles, the proposed methods were compared against existing baseline approaches and their performance were assessed.

Keywords: Feature Selection, Classification, Clustering, Regression, LSIR, PCA, MDS, Spectroscopy.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ANOVA	Análise de Variância
AUC	Área abaixo da curva
biPLS	*backward interval PLS*
IIV	Índice de importância de variáveis
iPLS	*interval PLS*
KNN	*k-Nearest Neighbors*
LDA	*Linear Discriminant Analysis*
LOOCV	*Leave-one-out cross-validation*
LR	Regressão Logística
LSIR	*Localized Slice Inverse Regression*
MDS	Escalonamento multidimensional
NB	*Naive Bayes*
NIR	*Near-infrared spectroscopy*
PA	Análise de Procrustes
PCA	*Principal Component Analysis*
PLS	*Partial Least Squares*
RMSE	Raiz quadrada do erro quadrático médio
RMSE	*Root Mean Square Error*
siPLS	*synergy interval PLS*
SIR	*Slice Inverse Regression*
SPA	*Successive Projections Algorithm*
SVM	*Support Vector Machine*
XRF	Espectroscopia por fluorescência de raios X

SUMÁRIO

1 Introdução

A espectroscopia reúne um conjunto de técnicas baseadas na absorção e emissão de luz e outras radiações pela matéria. Essas técnicas podem ser utilizadas para analisar amostras de diferentes naturezas quanto à sua composição química (GATIUS et al., 2017), e são amplamente utilizadas em diferentes áreas de estudo, como médica (GEYIK et al., 2021), alimentícia (WAFULA et al., 2022), bebidas (KAHMANN et al., 2017), farmacêutica (ANZANELLO et al., 2013), agricultura (XING et al., 2021) e petróleo (ALVES; POPPI, 2013). A popularidade de tais técnicas está associada à sua rapidez e simplicidade de execução, sendo uma técnica não destrutiva e que gera dados confiáveis (XIAOBO et al., 2010). No entanto, os instrumentos de análise por espectroscopia geralmente caracterizam uma amostra através da geração de centenas ou milhares de comprimentos de onda (variáveis). Assim, os bancos de dados espectroscópicos são tipicamente compostos por um elevado número de variáveis que, em muitos casos, excedem o número de observações e são altamente colineares (NG et al., 2019), os quais serão objeto de estudo nesta tese.

A presença de muitas variáveis redundantes nos bancos de dados espectroscópicos compromete a visualização dos dados e a eficácia dos algoritmos de aprendizado de máquina, bem como a compreensão das informações e a interpretação das análises (QU et al., 2019). Dessa forma, a aplicação de métodos de redução de dimensionalidade se torna necessária na calibração multivariada dos espectros, permitindo a identificação de padrões nos dados e a extração de informações relevantes (HUANG; LUO; XIA, 2019). A redução da dimensionalidade pode ser obtida através extração das variáveis, onde ocorre a transformação das variáveis originais dando origem a novas variáveis em um novo espaço de menor dimensão, ou através da seleção de variáveis, onde é identificado um subconjunto das variáveis originais mais informativas para o problema analisado (ZHOU et al., 2021).

Os métodos de extração das variáveis podem ser divididos em supervisionados e não-supervisionados, a depender da presença ou ausência de informações sobre as classes (ZHUO; CHENG; ZHANG, 2014). As técnicas não-supervisionadas, geralmente com propósitos exploratórios, são representadas por Análise de Componentes Principais (RENCHER, 2002), escalonamento multidimensional (BORG; P.J.F.; MAIR, 2017) e *Laplacian Eigenmaps* (SUNDARESAN; CHELLAPPA, 2008). Por sua vez, entre os algoritmos supervisionados têm-se *Fisher Linear Discriminant Analysis* (FDA) (BIAN;

TAO, 2014), *Neighbourhood Components Analysis* (NCA) (GOLDBERGER et al., 2005) e *Successively Orthogonal Discriminant Analysis* (SODA) (YU; MCKELVEY; KUNG, 2013), dentre outros tantos.

Os métodos de seleção de variáveis mais utilizados envolvem abordagens do tipo filtro, *wrapper* ou *embedded*. Os métodos do tipo filtro utilizam uma medida independente relacionada às características das variáveis para avaliá-las, sem envolver nenhum algoritmo de aprendizagem no processo (LI; LI; LIU, 2017). Como exemplo, pode-se citar os métodos baseados em similaridade como o ReliefF (ROBNIK-SIKONJA; KONONENKO, 2003), os métodos baseados em correlação, como o CFS (*Correlation-based feature selection*) (HALL, 1999), e os métodos baseados na teoria da informação, como o mRMR (*Maximum relevance and minimum redundancy*) (JU; HE, 2018). Os métodos do tipo *wrapper* utilizam algoritmos de aprendizado para avaliar os subconjuntos de variáveis e identificar os mais relevantes (BASGALUPP, 2007). Os algoritmos de otimização baseados em metaheurísticas como algoritmo genético (LEARDI, 2000), otimização por enxame de partículas (QASIM; ALGAMAL, 2018) e otimização da colônia de formigas (DORIGO; MANIEZZO; COLORNI, 1996) são característicos dessa abordagem. Por fim, os métodos do tipo *embedded* combinam os dois métodos anteriores, porém a seleção de variáveis e o aprendizado não podem ser separados (RODRIGUEZ-GALIANO et al., 2018), como executado pelos algoritmos *Random Forests* (RF) (HO, 1998), *Support Vector Machine – Recursive Feature Elimination* (SVM-RFE) (GUYON et al., 2002) e *Least Absolute Selection and Shrinkage Operator* (LASSO) (LEE; CAI, 2018). Destaca-se que os métodos do tipo *wrapper* e *embedded*, por estarem atrelados a algoritmos de aprendizado, apresentam um processo computacionalmente mais demorado do que os de filtro, além de estarem sob o risco de *overfitting* (ZHOU et al., 2021).

Reduzir a dimensionalidade de um conjunto de dados enquanto se preservam as informações relevantes ainda é um tópico ativo entre os pesquisadores. Abordagens com vistas à identificação das variáveis mais relevantes de bancos de espectroscópicos com fins de agrupamento, classificação ou predição de propriedades das amostras têm sido desenvolvidas em diferentes áreas de estudo, como no controle de qualidade do diesel (SOARES et al., 2017), autenticidade de medicamentos (ANZANELLO et al., 2017), identificação de espécies animais pelo sangue (ZHANG et al., 2021), detecção de defeitos internos em frutas (RAGHAVENDRA; GURU; RAO, 2021), classificação de amostras de

erva-mate de acordo com sua região de origem (KAHMANN et al., 2017), determinação de hemoglobina no sangue (TIAN et al., 2017) e predição de cafeína, trigonelina e ácido 5-cafeoilquínico em grãos de café (RIBEIRO; SALVA; SILVAROLLA, 2021).

Apesar da literatura abordar uma diversidade de métodos de redução de dimensionalidade, percebe-se que o uso de cada técnica pode ser influenciado pela natureza dos dados. De tal forma, desenvolver métodos apropriados e competitivos de seleção de comprimentos de onda que garantam a criação de modelos robustos relacionando comprimentos de onda (variáveis independentes) e propriedades de interesse (variáveis dependentes) ainda é um desafio para pesquisadores de diversas áreas. Assim, esta tese tem o intuito de desenvolver métodos inéditos de seleção de variáveis, com foco em dados espectrais, devido a sua alta utilidade, que possam ser utilizados com fins de agrupamento, classificação das amostras e predição de propriedades química, excluindo variáveis irrelevantes e ruidosas e construindo modelos mais precisos e fáceis de serem interpretados. Para avaliar o desempenho dos métodos propostos nesta pesquisa, os resultados obtidos são comparados com resultados de métodos tradicionais da literatura.

1.1 Tema e Objetivos

O tema desta tese contempla o uso de ferramentas multivariadas aplicadas à redução de dimensionalidade de bancos de dados altamente colineares. Como objetivo geral, tem-se o desenvolvimento de novas sistemáticas de seleção de variáveis espectrais com vistas ao agrupamento, classificação de amostras e predição de propriedades químicas, em bancos de dados de alta dimensionalidade.

Os seguintes objetivos específicos são elencados:

(i) Propor e validar um índice de importância de variáveis com vistas ao agrupamento de amostras de medicamentos;
(ii) Propor e validar um índice de importância de variáveis com vistas à regressão;
(iii) Demonstrar a importância da correta validação de modelos que incluem seleção de variáveis e a sua facilidade de reprodutibilidade;
(iv) Aperfeiçoar o método mRMR para possibilite capturar dependências lineares no cálculo de redundância de variáveis; e

(v) Avaliar o desempenho dos métodos propostos quando aplicados em diferentes bancos de dados e comparados a tradicionais métodos de seleção de variáveis da literatura.

1.2 Justificativa do tema e dos objetivos

As técnicas analíticas por espectroscopia estão sendo constantemente aprimoradas, tornando possível obter uma grande quantidade de informações sobre propriedades químicas de forma rápida e confiável. Entretanto, o enorme volume de dados gerados por tais técnicas pode conter dados irrelevantes e/ou ruidosos que acabam por limitar as capacidades preditivas ou exploratórias dos modelos gerados. Como resultado, o pré-tratamento dos dados coletados é um passo crítico na obtenção de modelos mais precisos. Considerando a seleção de variáveis como uma fase de pré-tratamento de dados, as sistemáticas sugeridas neste trabalho têm utilidade prática em termos de redução do número de variáveis utilizadas por algoritmos multivariados na previsão de qualidades químicas e classificação de produtos através de técnicas de espectroscopia.

Na atualidade, a manipulação e utilização de bancos de dados de grande volume, sejam eles pelo número de amostras ou pelo número de variáveis, é encontrado em praticamente todas as áreas do conhecimento que se valham de abordagens quantitativas. Por mais que do ponto de vista teórico existam diversos métodos de seleção de variáveis que possam retornar uma solução ótima, todos os métodos se valem de suposições que muitas vezes podem ser reais em algum campo do conhecimento, porém não em outros. Portanto, o desenvolvimento de abordagens para selecionar as variáveis mais relevantes na geração de modelos de previsão e análise confiáveis e parcimoniosas continua sendo um importante nicho na literatura acadêmica, dado que pesquisadores ou tomadores de decisão podem preferir modelos mais enxutos ao custo de pequenos decréscimos de capacidade preditiva. Além disso, tais modelos podem ser mais interpretáveis, ou possibilitar a miniaturização ou simplificação de complexos aparatos experimentais. Como resultado, o estudo da combinação de estratégias multivariadas e sistemáticas para selecionar as variáveis mais relevantes para a construção de modelos de classificação e regressão mais robustos, baseados em técnicas analíticas, justifica esta pesquisa.

1.3 Delineamento do Estudo

Com intuito de delinear os meios através dos quais os objetivos propostos serão alcançados, essa seção apresenta o método de pesquisa, o método de trabalho utilizado, assim como a estrutura da tese, com um resumo das ferramentas utilizadas e contribuições científicas de cada artigo que compõe a tese.

1.3.1 Método de Pesquisa

Quanto à natureza, a presente tese é classificada como pesquisa aplicada, visto que a fundamentação teórica é explorada e direcionada à solução de problemas genéricos (CRESWELL, 2010). Em relação aos objetivos, essa pesquisa é enquadrada como pesquisa exploratória, uma vez que proporciona o conhecimento e visão geral do problema, possibilitando a construção de hipóteses precisas e operacionalizáveis para solucioná-lo (GIL, 2008). Por fim, essa pesquisa apresenta uma abordagem quantitativa, utilizando análises estatísticas e modelagem matemática para encontrar as soluções dos problemas apresentados, fazendo uso de análises numéricas e propiciando análises estatísticas da realidade (BERTO; NAKANO, 1999).

1.3.2 Método de Trabalho

O desenvolvimento da pesquisa apresentada nessa tese ocorre através de quatro etapas, cada uma delas corresponde a um artigo com o intuito de atender os objetivos da tese.

No primeiro artigo é proposto um método não supervisionado para seleção de variáveis com foco no agrupamento de amostras de medicamentos. O método é baseado na integração de técnicas multivariadas de escalonamento multidimensional (*Multidimensional Scaling* - MDS) e análise de Procrustes (PA). O MDS fornece uma projeção dos dados originais em um subespaço reduzido buscando manter a relação de distâncias entre amostras, enquanto que a análise de Procrustes encontra uma matriz de projeção a partir dos dados originais nesse novo subespaço. Os pesos da matriz resultante da análise de Procrustes dão origem a um índice de importância de variáveis que orienta um processo iterativo de seleção após cada variável ser inserida no modelo. Na sequência, é conduzido um procedimento de

otimização baseado em um algoritmo de busca para maximizar o desempenho de clusterização, o qual é avaliado através do Índice de Silhouette (SI).

No segundo artigo, uma abordagem para seleção de variáveis baseada nos pesos da matriz de projeção da regressão inversa em fatias localizadas (*Localized Sliced Inverse Regression* - LSIR) é apresentada. Os pesos da matriz de projeção são usados como uma proxy para a importância de variáveis e, similarmente ao artigo 1, são a base do índice que guia o processo de seleção de variáveis que visa minimizar tanto o erro quadrático médio na partição de validação, quanto o número de comprimentos de onda retidos.

O terceiro artigo apresenta uma detalhada tentativa de replicação de um recente estudo que emprega a espectroscopia Raman do soro de sangue para classificação entre pacientes com COVID-19 e pacientes controle. O artigo demonstra duas principais falhas metodológicas em sua versão original: a não inclusão do processo de seleção de variáveis dentro do *loop* de validação, e a ambiguidade da descrição do método, o que tornou impossível a replicação dos resultados dado o mesmo banco de dados. Além disso, o artigo serve como motivação para o quarto artigo.

O quarto artigo tem como objetivo a melhoria do desempenho de classificação de pacientes positivos ou negativos para COVID. Para tanto, é utilizado o mesmo banco de dados do artigo 3, porém é proposta uma melhoria no tradicional método de seleção de variáveis mRMR (máxima relevância mínima redundância). O método proposto, PCA-mRMR, relaxa uma das suposições do mRMR tradicional, o qual assume que as variáveis selecionadas em cada passo são independentes das demais. Ao aplicar PCA a cada interação do processo de seleção de variáveis, o método proposto busca capturar as relações lineares entre as variáveis já selecionadas, enquanto o mRMR tradicional apenas leva em conta comparações par-a-par. Os resultados são comparados com baselines já existentes.

1.3.3 Estrutura da tese

A estrutura da tese corresponde ao desenvolvimento de quatro artigos. Na Tabela 1.1 são apresentados os artigos e a comunicação que compõem a tese, ferramentas utilizadas, e as contribuições científicas de cada artigo.

Tabela 1.1. Descrição dos artigos que compõem esta tese.

Estudos	Título	Ferramentas utilizadas	Contribuição científica
Artigo 1 [(a)]	Enhancing counterfeit and illicit medicines grouping via feature selection and X-ray fluorescence (XRF) spectrometry	Escalonamento multidimensional, Análise de Procrustes	a) Novo método de seleção de variáveis para o agrupamento de medicamentos em grupos de originais ou falsificados; b) Utilização do escalonamento multidimensional integrado à análise de Procrustes para a geração de um novo índice de importância de variáveis.
Artigo 2 [(b)]	A wavelength selection method based on Localized Sliced Inverse Regression weights	Localized sliced inverse regression (LSIR), Partial Least Square regression (PLS)	a) Novo índice de importância de variáveis baseado nos pesos gerados pelo LSIR.
Artigo 3 [(c)]	Communication regarding the article "An efficient primary screening COVID-19 by serum Raman spectroscopy"	Análise Discriminante, ANOVA, Support Vector Machine (SVM)	a) Demonstração da importância da divisão do banco de dados em conjunto de treinamento e teste de forma adequada.
Artigo 4 [(d)]	Improved Maximum Relevancy Minimum Redundancy algorithm for high order linear dependencies: application to COVID-19 screening with Raman spectroscopy	Análise de Componentes Principais (PCA) Support Vector Machine (SVM)	a) Aprimoramento do método de seleção de máxima relevância mínima redundância (mRMR)

(a) Artigo publicado no periódico Journal of Pharmaceutical and Biomedical Analysis
(b) Artigo a ser submetido ao periódico Chemometrics and Intelligent Laboratory Systems
(c) Artigo publicado no periódico Journal of Raman Spectroscopy
(d) Artigo a ser submetido ao periódico Journal of Raman Spectroscopy

1.4 Delimitações do Estudo

A presente pesquisa se concentra na proposição de métodos de seleção de variáveis, visando a redução de dimensionalidade dos bancos de dados. Todas as ferramentas utilizadas são oriundas da literatura, de forma que não serão desenvolvidas novas ferramentas de análise de dados. Para o agrupamento das amostras foi utilizado o método hierárquico aglomerativo; para a regressão com foco na predição de propriedades químicas das amostras foi utilizada a regressão PLS; e, para a classificação das amostras, os modelos SVM foram construídos utilizando apenas kernel RBF. A presente tese teve como foco apenas relações lineares entre as variáveis sendo analisadas, dado que este é um dos principais cenários encontrados, enquanto SVM com kernel RBF foi utilizado somente para replicar o mesmo cenário experimental do experimento de base.

Como métricas de avaliação da performance das ferramentas de agrupamento, predição e classificação foram utilizadas o *Silhouette Index*, raiz do erro quadrático médio, acurácia, especificidade e sensibilidade. Em relação à abrangência, a pesquisa utilizou bancos de dados nas áreas farmacêutica, biologia, alimentícia, petroquímica e médica.

1.5 Estrutura da Tese

Esta proposta de tese está organizada em seis capítulos. No primeiro capítulo foram expostos a contextualização da pesquisa, apresentando o tema, objetivos e justificativas da tese, bem como o método de trabalho adotado e as delimitações do estudo. Os capítulos 2, 3, 4, e 5 apresentam os artigos desenvolvidos na tese, seguindo a estrutura apresentada anteriormente. O sexto e último capítulo traz as considerações finais desta tese de doutorado, bem como sugestões para trabalhos futuros.

1.6 Referências

AKIN GEYIK, G. et al. A rapid diagnostic approach for gastric and colon cancers via Fourier transform mid–infrared spectroscopy coupled with chemometrics from paraffin–embedded tissues. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, p. 120619, nov. 2021.

ALVES, J. C. L.; POPPI, R. J. Biodiesel content determination in diesel fuel blends using near infrared (NIR) spectroscopy and support vector machines (SVM). **Talanta**, v. 104, p. 155–161, 2013.

ANZANELLO, M. J. et al. A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes. **Journal of Pharmaceutical and Biomedical Analysis**, v. 83, p. 209–214, 2013.

ANZANELLO, M. J. et al. A genetic algorithm-based framework for wavelength selection on sample categorization. **Drug Testing and Analysis**, v. 9, n. 8, 2017.

BASGALUPP, M. P. **Algoritmos genéticos para seleção de atributos em problemas de classificação de processos de negócio**. [s.l.] PUC-RS, 2007.

BERTO, R. M. V. S.; NAKANO, D. N. A produção científica nos anais do encontro nacional de engenharia de produção: um levantamento de métodos e tipos de pesquisa. **Production**, v. 9, n. 2, p. 65–75, dez. 1999.

BIAN, W.; TAO, D. Asymptotic Generalization Bound of Fisher's Linear Discriminant Analysis. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 36, n. 12, p. 2325–2337, 1 dez. 2014.

BORG, I.; P.J.F., G.; MAIR, P. **Applied multidimensional scaling and unfolding**. [s.l.] Springer, 2017.

CRESWELL, J. W. **Projeto de pesquisa: métodos qualitativo, quantitativo e misto**. [s.l.] Artmed, 2010.

DORIGO, M.; MANIEZZO, V.; COLORNI, A. Ant system: optimization by a colony of cooperating agents. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, v. 26, n. 1, p. 29–41, fev. 1996.

GATIUS, F. et al. Comparison of CCA and PLS to explore and model NIR data. **Chemometrics and Intelligent Laboratory Systems**, v. 164, n. March, p. 76–82, 2017.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6ª ed. São Paulo: Atlas S. A., 2008.

GOLDBERGER, J. et al. Neighbourhood Components Analysis. **Advances in Neural Information Processing Systems**, v. 17, p. 513–520, 2005.

GUYON, I. et al. Gene Selection for Cancer Classification using Support Vector Machines. **Machine Learning**, v. 46, p. 389–422, 2002.

HALL, M. A. **Correlation-based feature Selection for Machine Learning**. [s.l.] University of Waikato, 1999.

HUANG, X.; LUO, Y.-P.; XIA, L. An efficient wavelength selection method based on the maximal information coefficient for multivariate spectral calibration. **Chemometrics and Intelligent Laboratory Systems**, v. 194, p. 103872, nov. 2019.

JU, Z.; HE, J. J. Prediction of lysine glutarylation sites by maximum relevance minimum redundancy feature selection. **Analytical Biochemistry**, v. 550, n. January, p. 1–7, 2018.

KAHMANN, A. et al. Near infrared spectroscopy and element concentration analysis for assessing yerba mate (Ilex paraguariensis) samples according to the country of origin. **Computers and Electronics in Agriculture**, v. 140, p. 348–360, 2017.

LEARDI, R. Application of genetic algorithm-PLS for feature selection in spectral data sets. **Journal of Chemometrics**, v. 14, n. 5–6, p. 643–655, 2000.

LEE, C.-Y.; CAI, J.-Y. LASSO variable selection in data envelopment analysis with small datasets. **Omega**, dez. 2018.

LI, Y.; LI, T.; LIU, H. Recent advances in feature selection and its applications. **Knowledge and Information Systems**, v. 53, n. 3, p. 551–577, 5 dez. 2017.

NG, W. et al. Optimizing wavelength selection by using informative vectors for parsimonious infrared spectra modelling. **Computers and Electronics in Agriculture**, v. 158, n. February, p. 201–210, 2019.

QASIM, O. S.; ALGAMAL, Z. Y. Feature selection using particle swarm optimization-based logistic regression model. **Chemometrics and Intelligent Laboratory Systems**, v. 182, p. 41–46, nov. 2018.

QU, Y. et al. Non-unique decision differential entropy-based feature selection. **Neurocomputing**, jul. 2019.

RAGHAVENDRA, A.; GURU, D. S.; RAO, M. K. Mango internal defect detection based on optimal wavelength selection method using NIR spectroscopy. **Artificial Intelligence in Agriculture**, v. 5, p. 43–51, 2021.

RENCHER, A. C. **Methods of Multivariate Analysis**. New York: Wiley Interscience, 2002.

RIBEIRO, J. S.; SALVA, T. DE J. G.; SILVAROLLA, M. B. Prediction of a wide range of compounds concentration in raw coffee beans using NIRS, PLS and variable selection. **Food Control**, v. 125, p. 107967, jul. 2021.

ROBNIK-SIKONJA, M.; KONONENKO, I. Theoretical and empirical analysis of {R}elief and {R}elief{F}. **Machine Learning**, v. 53, p. 23–69, 2003.

RODRIGUEZ-GALIANO, V. F. et al. Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. **Science of The Total Environment**, v. 624, p. 661–672, maio 2018.

SOARES, F. et al. A non-equidistant wavenumber interval selection approach for classifying

diesel/biodiesel samples. **Chemometrics and Intelligent Laboratory Systems**, v. 167, n. June, p. 171–178, 2017.

SUNDARESAN, A.; CHELLAPPA, R. Model Driven Segmentation of Articulating Humans in Laplacian Eigenspace. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 30, n. 10, p. 1771–1785, out. 2008.

TIAN, H. et al. Optical wavelength selection for portable hemoglobin determination by near-infrared spectroscopy method. **Infrared Physics & Technology**, v. 86, p. 98–102, nov. 2017.

TIN KAM HO. The random subspace method for constructing decision forests. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 20, n. 8, p. 832–844, 1998.

WAFULA, E. N. et al. Antinutrient to mineral molar ratios of raw common beans and their rapid prediction using near-infrared spectroscopy. **Food Chemistry**, v. 368, p. 130773, jan. 2022.

XIAOBO, Z. et al. Variables selection methods in near-infrared spectroscopy. **Analytica Chimica Acta**, v. 667, n. 1–2, p. 14–32, 2010.

XING, Z. et al. A method combining FTIR-ATR and Raman spectroscopy to determine soil organic matter: Improvement of prediction accuracy using competitive adaptive reweighted sampling (CARS). **Computers and Electronics in Agriculture**, v. 191, p. 106549, dez. 2021.

YU, Y.; MCKELVEY, T.; KUNG, S.-Y. **A classification scheme for high-dimensional-small-sample-size data using soda and ridge-SVM with microwave measurement applications**. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. **Anais**...IEEE, maio 2013Disponível em: <http://ieeexplore.ieee.org/document/6638317/>

ZHANG, L. et al. Optimal wavelengths selection from all points for blood species identification based on spatially resolved near-infrared diffuse transmission spectroscopy. **Infrared Physics & Technology**, v. 117, p. 103865, set. 2021.

ZHOU, R. et al. Supervised Dimensionality Reduction Technology of Generalized Discriminant Component Analysis and Its Kernelization Forms. **Pattern Recognition**, p. 108450, nov. 2021.

ZHUO, L.; CHENG, B.; ZHANG, J. A comparative study of dimensionality reduction methods for large-scale image retrieval. **Neurocomputing**, v. 141, p. 202–210, out. 2014.

2 ARTIGO 1 – Enhancing counterfeit and illicit medicines grouping via feature selection and X-ray fluorescence (XRF) spectrometry

Published in Journal of Pharmaceutical and Biomedical Analysis, v. 174, 2019.

Abstract

In this paper, we propose a novel framework to select the most relevant X-Ray Fluorescence (XRF) energy values (i.e., features) to enhance the clustering (grouping) of counterfeit and illicit medical tablets. The framework is based on the integration of multidimensional scaling (MDS) and Procrustes analysis (PA) multivariate techniques. MDS provides a projection of the original data into a lower dimension, while PA finds a projection matrix from the original data. Such outputs give rise to a feature importance index that guides an iterative feature selection process; after each feature is inserted in the subset, an optimization procedure based on a greedy search method is carried out to maximize the clustering quality assessed through the Silhouette Index (SI). The inorganic chemical fingerprinting of 41 commercial samples (Viagra®, Cialis®, Lazar®, Libiden®, Maxfil®, Plenovit®, Potent 75®, Rigix®, V-50®, Vimax® and Pramil®) and 56 seized counterfeit samples (Viagra and Cialis) was used to validate the proposed framework. From the original 2048 data points in the full spectra, we identified a subset comprised of 41 energy values that substantially improved clustering quality; the obtained groups were assessed by visual inspection of the PCA plots.

Keywords: XRF Spectrometry, Feature Selection, Counterfeit Medicines, PCA.

2.1 Introduction

The commerce of counterfeit and illicit medicines is an international issue that has considerably increased since many years. Since the production of counterfeit medicines do not follow strict manufacturing processes and quality assessment, they pose a serious risk to public health [1]. One of the most seized medicines in industrialized countries and "new economies countries" are the phosphodiesterase type 5 (PDE-5) inhibitors, which are used to treat erectile dysfunction [2]. These drugs are of special interest given their high cost and potential embarrassment associated with the underlying condition, leading customers to rely on internet websites as main buying sources [3].

In Brazil, Viagra® and Cialis® are the most counterfeited medicines, accounting for a substantial number of seizures carried out by the Brazilian Federal Police (BFP) [4]. In addition to seizure operations, the BFP also performs forensic analyses on the counterfeit tablets in an attempt to combat this increasing phenomenon [4]. Being able to fingerprint counterfeit medicines and associate different seized batches to a common fraudulent source is a key aspect of the forensic analysis, since it can be useful to unveil sophisticated counterfeiting criminal operations. That can be treated as a clustering problem [5] from a computational perspective as performed in several related works relying on different products (e.g., beverages adulteration [6], counterfeit medicines in general [7]), and analytical techniques (Raman [8] and vibrational [9] spectroscopy).

Among the many available chemical techniques employed to sample analysis, X-Ray Fluorescence (XRF) is the one suitable for detecting and characterizing the presence of metals in samples [10,11]. Some of the advantages of XRF rely on its multielemental capability, high precision, short analysis times and good detection, among others. In light of that, XRF is deemed an important analytical method for the determination of active ingredients, excipients, and covering agents, such as calcium phosphate, titanium oxide, and iron oxide [4] in medicines. On the other hand, the size of the energy spectra generated by XRF may jeopardize the performance of several statistical and multivariate tools applied to such spectra with the purpose of sample characterization, requiring efforts towards the identification of the most informative regions of the spectra.

In this paper, we employ XRF data in conjunction with a novel feature selection framework based on multidimensional scaling and Procrustes analysis to perform the inorganic fingerprinting of several authentic (i.e., Viagra®, Cialis®, Lazar®, Libiden®, Maxfil®, Plenovit®, Potent 75®, Rigix®, V-50®, Vimax®, and Pramil®) and unauthentic medicines indicated for erectile dysfunction. While XRF is capable of capturing detailed information regarding metal presence in the samples [10,11], the proposed framework for feature selection enables the identification of the most relevant energy values that can improve clustering results and possibly unveil additional information not present when the whole spectrum is assessed. Such clusters can provide investigative forces with relevant information towards tracking and interrupting counterfeit operations (e.g. samples from different seizures inserted into the same cluster may suggest a common illegal source).

2.2 Material and methods

In this section, we describe the assessed samples, the multivariate techniques used, and the proposed framework for feature selection.

2.2.1 Samples

The assessed dataset is comprised of 97 samples of original and counterfeit medicines for erectile dysfunction; each sample is described by 2048 energy values (i.e., features). Eight authentic samples of Viagra® and Cialis®, containing 50 mg of sildenafil and 20 mg of tadalafil, were supplied by Pfizer Ltda and Eil Lilly do Brasil Ltda laboratories, respectively. In addition, 33 deemed as authentic tablets containing sildenafil citrate and/or tadalafil of several trademarks were supplied by the BFP, as well as 56 counterfeit samples from different seizures. Details about the dataset are given in Table 2.1.

Table 2.1: Description of samples and seizures for the assessed tablets.

Tablets	Number of samples	Register of seizure
Authentic Viagra®	4	-
Counterfeit Viagra	6	I
	10	II
	1	III
	1	IV
Authentic Cialis®	4	-
Counterfeit Cialis	10	V
	10	VI
	18	I
Lazar®	3	-
Potent 75®	2	-
Vimax®	3	-
Maxfil®	3	-
Rigix®	2	-
Plenovit®	6	-
Pramil®	10	-
V-50®	3	-
Libiden®	1	-

The ED-XRF (Energy Dispersive-XRF) spectra were acquired using a EDX-700 (Shimadzu, Kyoto, Japan) spectrometer. Measurements were performed under air, with a beam collimation of 3 mm, 25% of detector dead time, with current automatically adjusted during acquisition to keep the detector dead time of 25%. For spectral acquisition, tablets were crushed using a mortar and placed into XRF cells on Mylar™ film (3 μm thickness). The measurement time was 250 s, with spectra recorded from 0 to 40 keV, with energy step

of 0.02 keV, leading to 2048 points (i.e., features) for each spectrum. Figure 2.1 shows the resulting spectra of the four main classes: Authentic Viagra, Counterfeit Viagra, Authentic Cialis, and Counterfeit Cialis.

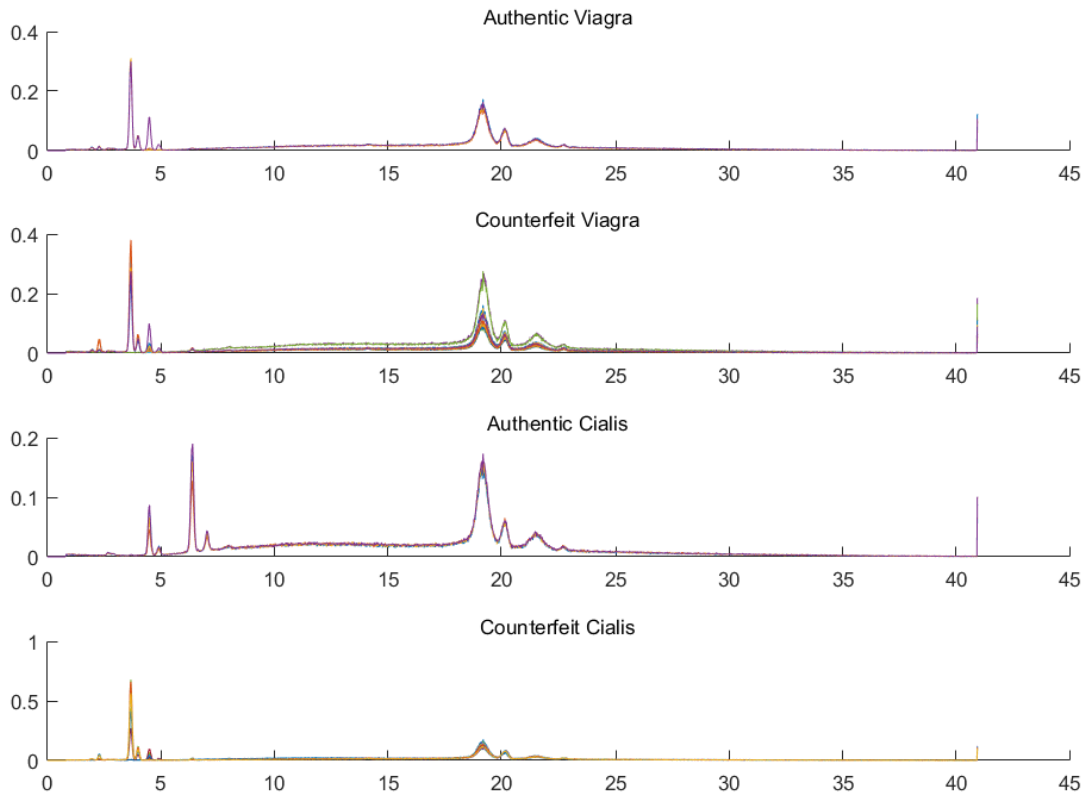

Figure 2.1: ED-XRF spectra of authentic and counterfeit samples of Viagra and Cialis.

2.2.2 Multivariate techniques

There are two multivariate techniques employed in our propositions, the Multidimensional scaling (MDS) and the Procrustes Analysis (PA). The MDS, also known as Principal Coordinates Analysis, is a visualization technique that works by reducing the dimensionality of the data [12]. While in Principal Component Analysis (PCA) the aim is to project the original data points into a lower dimensional space preserving the maximum variance, in MDS the objective is to represent the data points into an *N*-dimensional space that preserves the distances between the assessed objects (i.e. samples) [13].

The classical MDS, which will be used in this work, considers a distance matrix of size $n \times n$, where n is the number of samples, for the original data points as input and finds a new coordinate matrix by eigenvalue decomposition. The steps for this procedure are as follows [14]: (i) Derive a **D** squared-distance matrix ($n \times n$) from the sample-feature matrix **X** between data points (i.e., samples) i and j using the Euclidean distance: $\mathbf{D}_{ij} = \left(\mathbf{X}_i - \mathbf{X}_j\right)^{\mathrm{T}}(\mathbf{X}_i - \mathbf{X}_j)$; (ii) Double center the **D** matrix, originating $\mathbf{B} = -\frac{\mathbf{1}}{\mathbf{2}}\mathbf{LDL}$, where $\mathbf{L} = \mathbf{I} - \frac{1}{\mathrm{n}}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ is the centering matrix, n is the total number of samples, **I** is the identity matrix and 1 is a vector of ones. At this step, the new **B** matrix is column and row centered; (iii) Given that **B** is symmetric positive semi-definite, it can be decomposed as $\mathbf{B} = \mathbf{V}\Delta\mathbf{V}^{\mathbf{T}}$, where $\Delta = diag(\lambda_1, \ldots, \lambda_n)$ is the diagonal matrix of eigenvalues of **B**, and $\mathbf{V} = [\mathrm{V}_1, \ldots, \mathrm{V}_n]$ is the corresponding matrix of eigenvectors; and (iv) Taking the first p largest eigenvalues, and their corresponding eigenvectors, the new coordinate matrix **Y** can be computed as $\mathbf{Y} = \mathbf{V}_{\mathrm{p}}\boldsymbol{\Delta}^{1/2}$. For visualization, p usually equals to 2.

As for the Procrustes Analysis (PA), it aims at determining a linear transformation comprised of translation, reflection, orthogonal rotation, and scaling operations, of the data points in matrix **Z** aimed at maximizing their matching with the points in another **Y** matrix [15]. This is carried out by minimizing the root mean square distance between the set of data points in the two matrices. The **Z** matrix can be expressed as $\mathbf{Z} = b\mathbf{YT} + c$, where b is the scale component, **T** is the reflection and rotation component, and c is the translation component. Details about the operationalization of PA are given in [16,17].

2.2.3 Proposed method for feature selection

It has been shown in related works that the weights derived from PCA [18–20] or Partial Least Squares (PLS) [21–23] can be used as an indication of feature importance. The rationale is that if a given feature *f* has a large absolute coefficient in the projection matrix, this feature offers a larger impact to the projected subspace when compared to remaining features, thus being deemed more important. Inspired by those findings, our propositions integrate MDS with PA to derive a new unsupervised feature importance index.

Despite of being conceptually related to PCA, MDS does not give rise to a projection matrix to derive importance indices, as verified for the latter. To overcome this fact, we apply PA to the matrix resulting from MDS and the original matrix **X**, and then derive an

importance index from the reflection and rotation components from the PA. Given a mean centered and unit-variance matrix **X** describing the XRF data, the proposed framework is operationalized in 7 steps:

Step 1 - Compute the distance matrix **D** (see equation in Section 2.2.2) by using the correlation distance; we recommend that distance since it is more suitable for highly correlated features, such as XRF data;

Step 2 - Apply the classical MDS to **D** and store the first p eigenvectors in **Y**; p can be defined by visual inspection of the ordered eigenvalues plot, similarly to what is usually performed when determining the best number of principal components in PCA [24];

Step 3 - Perform a PA of **X** and **Y** and retrieve the reflection and rotation component **T**, which is related to the original features in **X**;

Step 4 - Normalize **T** by mean centering and scaling it to unit-variance. This is carried out to capture the relative importance among the F features;

Step 5 - The importance index $W_f (f=1,\ldots,F)$ is evaluated for each feature f by summing their absolute values for each component of **T**;

Step 6 - Features are ordered according to the decreasing value of w_f, which denotes the relevance of feature f;

Step 7 - Once features are ordered according to their importance, the best number of features is defined by optimizing a given clustering quality metric. In our framework, we suggest the use of a greedy search method to optimize the Silhouette Index (SI) [25] by varying the number of features included in the model, which are added according to the importance index, and the number of clusters. As for the clustering method, we employ the agglomerative hierarchical algorithm with single linkage, since it does not depend on initialization parameters [26].

2.3 Results and Discussion

We now present the results of our propositions applied to the dataset described in Section 2.1.

2.3.1 Feature ranking and selection

Features (energy values) were ranked according to their importance as proposed in Section 2.3. We defined p=2 in step 2, since the first two eigenvalues from MDS are almost 15 and 6 times larger than the third one, respectively, as depicted in Figure 2.2.

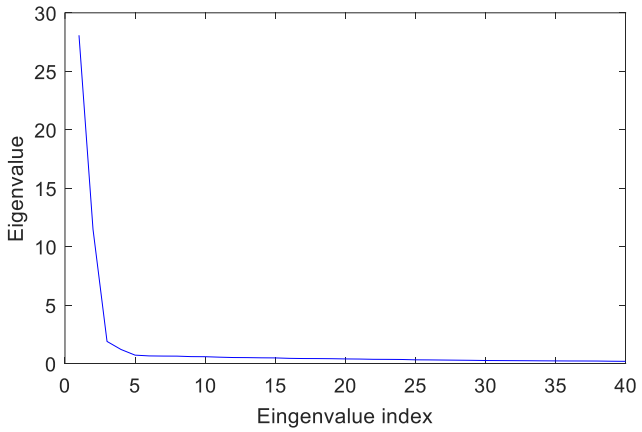

Figure 2.2: First forty eigenvalues from MDS.

After evaluating the importance index w for all F features, we plotted the resulting weights according to the energy values to have an overall visualization for the entire set of features (see Figure 2.3). One can notice that the features deemed more important are related to energy values around 1keV to 4keV, as well as for energies within the 5keV to 10keV interval.

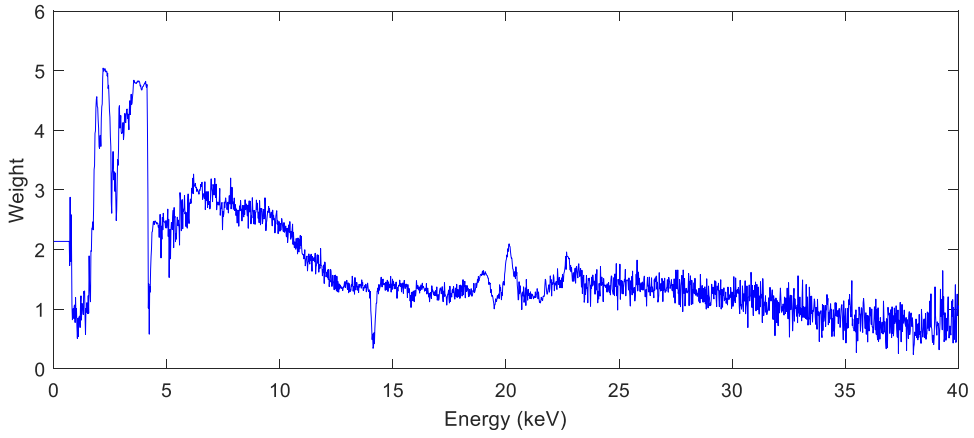

Figure 2.3: Importance weights according to energy.

We then proceeded to the greedy search step which relies on two main purposes: (i) define the best number of clusters to be formed; and (ii) determine the most informative features (i.e. energy) for cluster formation. We set an upper bound of 200 features, while the number of clusters ranged from 4 to 15 since a preliminary study with the same dataset [4] identified at least 6 possible clusters. In Figure 2.4a we show the complete SI surface relating number of clusters and retained features; the best configuration is indicated by the arrow in Figure 2.4b, when 41 out of the 2048 original features are used to generate 5 clusters.

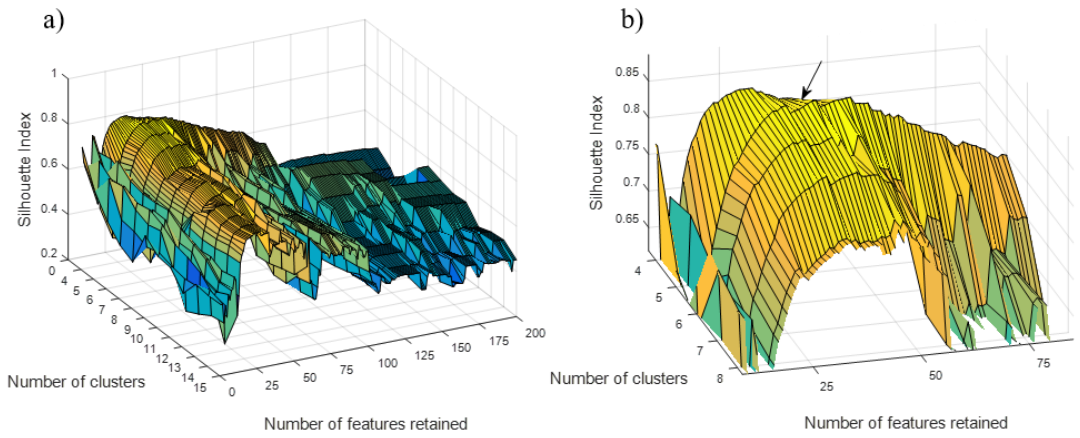

Figure 2.4: Silhouette Index for different number of features and clusters. Plot a) shows the complete surface, while in b) we detail the region presenting the highest SI values; the maximum SI is identified by the arrow.

We then assessed the spectral regions corresponding to the retained features, which are highlighted in Figure 2.5 a); the importance indices *w* for those features are presented in Figure 2.5 b). One can notice that the regions from 2.10 keV to 2.40 keV, from 3.50 keV to 3.80 keV, and from 3.90 keV to 4.10 keV are the ones selected by the proposed framework. Such regions show the presence of Kα characteristic lines for P (1.98 keV), Ca (3.62 keV), and Kβ for Ca (4.0 keV), which are found in authentic Viagra [4] and are related to the excipient dibasic calcium phosphate ($CaHPO_4$); that is not found in Cialis samples. Regarding Cialis, the characteristic lines of Ti and Fe, related to the excipient titanium dioxide (TiO_2) and covering agent [4], respectively, were not retained (Kα (4.48 keV) and Kβ (4.9 keV) for Ti, Kα (6.38 keV) and Kβ (7.02 keV) for Fe). This is of special interest, since they are mainly related to Cialis, contributing to a better discrimination of such samples.

Another interesting finding is that the Kα characteristic line for S (3.30 keV) is not retained, possibly due to the fact that sulfur is present in the sildenafil citrate structure, which is the active ingredient of Viagra[4]. In addition, as shown in a previous work [27] some counterfeit Cialis Samples contained contents of sildenafil citrate, and thus sulfur (i.e. the same powder mixture was being used to produce counterfeit Viagra and Cialis). Since no placebo samples were found in the assessed seizures, S did not improve the sample stratification. Also, as the excipient represents the largest amount of material in a tablet, differences in excipient's composition (e.g. lactose) may be more discriminant than the active compound itself.

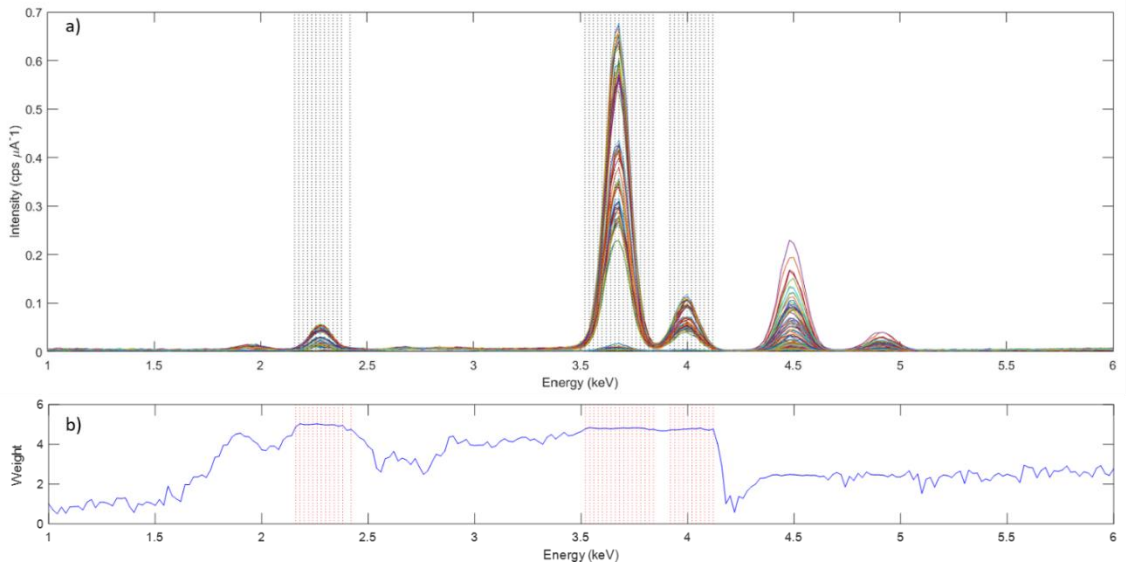

Figure 2.5: a) XRF spectra with the selected energy values indicated by vertical dashed lines, and b) importance index for the energy values, with the selected ones indicated by vertical dashed lines.

In Figure 2.6, we show the individual plots for each of the 13 subclasses studied. As aforementioned, none of the selected regions are encountered in authentic Cialis, which is seen as an appropriate discriminatory feature. In addition, it is noteworthy that two samples of Pramil and Plenovit present high intensity values in the region 3.50 keV to 3.80 keV, probably associated to outlier samples. In the additional material Figure S2.1, we show the PCA plots of such samples showing their unlike profile to the rest of the samples of the same class. This can be due to the limitation of XRF regarding particle granularity, which can hinder the spectra acquisition and lead to high standard deviations. Libiden also presents a similar profile to authentic Cialis in the selected regions, thus both medicines should be grouped together.

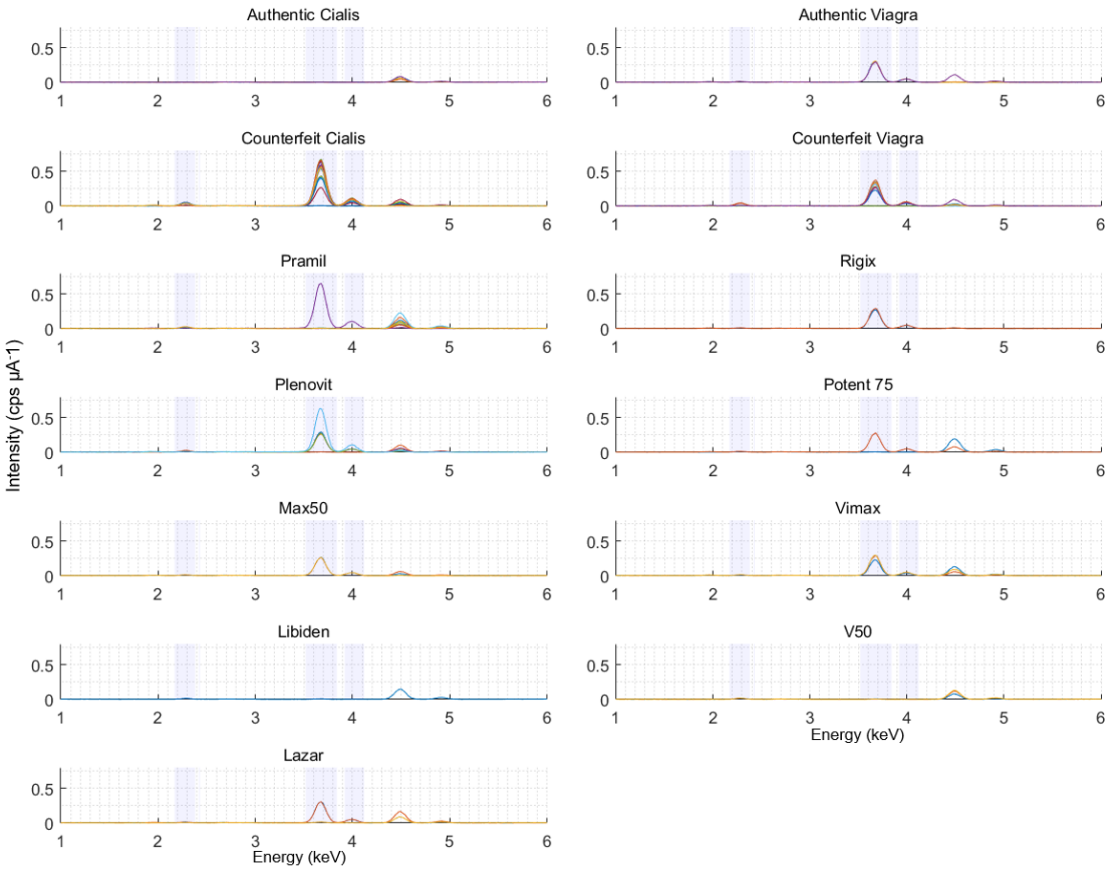

Figure 2.6: XRF spectra of all 13 subclasses analyzed; selected energy values are indicated by the shaded regions.

2.3.2 Visualization and comparison

PCA was then applied to the 41 features identified as the most informative by the proposed framework. Figure 2.7 a) shows the first two principal components, with samples identified according to the source. Such subset of features is deemed relevant for the stratification as almost all counterfeit samples are isolated from genuine ones, with a few exceptions. At least seven clusters can be identified in this plot, one of them mixing both original and counterfeit tablets.

Aimed at visually comparing the results of our propositions, we also graphed the 2 PCs derived from the original 2048 features, which is shown in Figure 2.7 b). The most important difference is that the use of all features masks one of the counterfeit Cialis clusters, which is very distinguishable when our framework is applied. However, two counterfeit Viagra samples, which are single samples accounting for seizures III and IV, were included in the cluster with most of the samples pertaining to authentic medicines of different

suppliers. After analyzing the individual spectra, we identified that both samples do not present peaks in the selected intervals, suggesting their composition to be predominantly organic [4]. In addition, two authentic samples from Plenovit® and Pramil® with high intensities around 3.50 keV to 3.80 keV (depicted in Figure 2.6) are included in the counterfeit Cialis cluster in both plots, which may indicate outliers or a poor quality control process, especially given that samples from these two medicines can be found in more than one cluster. Other possible reason, which has already been studied in the previous work [27] is that both samples are not authentic and were manufactured using the same powder. Since the samples were not supplied directly by the manufacturer, we cannot assume that all are truly original samples. We can also disregard pure experimental error, since the samples are well grouped with the counterfeit Cialis group in both full spectra and the selected one. If experimental error would be the cause of such discrepancy, the samples would probably not be that well clustered, but rather scattered around. In the additional material, Figure S2.2, we show the PCA plot of the Plenovit® samples.

The comparison of the PCA plot using the 41 selected features out of 2048 (2%) with the plot relying on all features corroborates the efficiency and robustness of our method. We were able to unveil an additional cluster comprised of counterfeit Cialis samples, while also reducing the number of features needed for group formation. In addition, the selected features are consistent with the common excipients used in the medications.

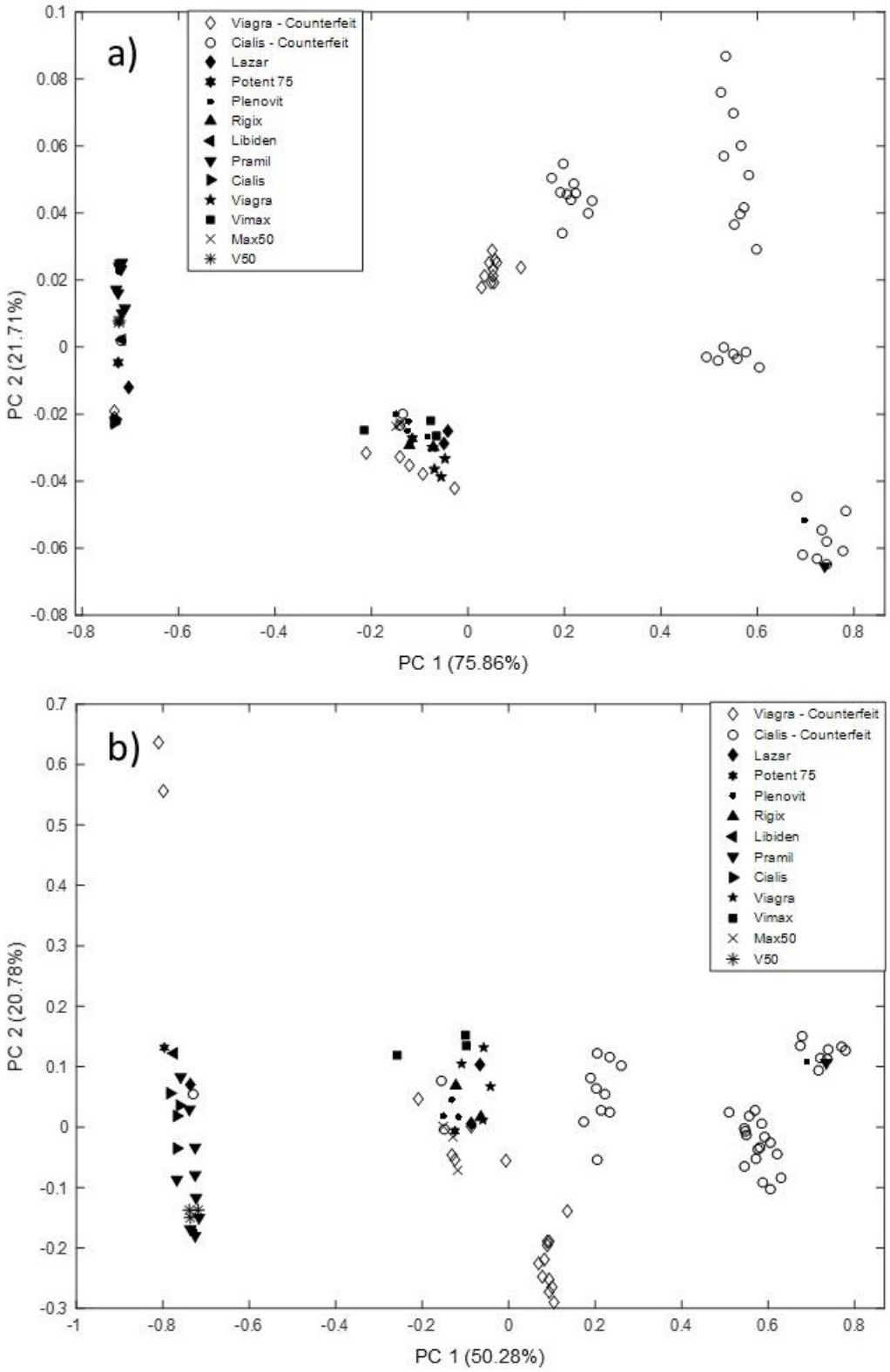

Figure 2.7: Plots of the first two principal components of PCA considering a) the 41 selected features, and b) all spectra.

2.3.3 Comparison to another wavelenght selection method

For the sake of comparison with another method of wavelength selection, we decided to implement and use the same method as described in [20], which uses the PCA scores to derive a feature importance index. The absolute value of the scores of each feature are deemed to relate to the feature importance.

Similarly to our method, we used PCA to create an importance index and then proceeded with the same steps 6 and 7 of our method, which is the responsible by determining the correct final number of features. To determine the number of PCA components, we performed an analysis of the explained variance in each component. The final number of PCA components was set to 2, which can be confirmed in Figure 2.8.

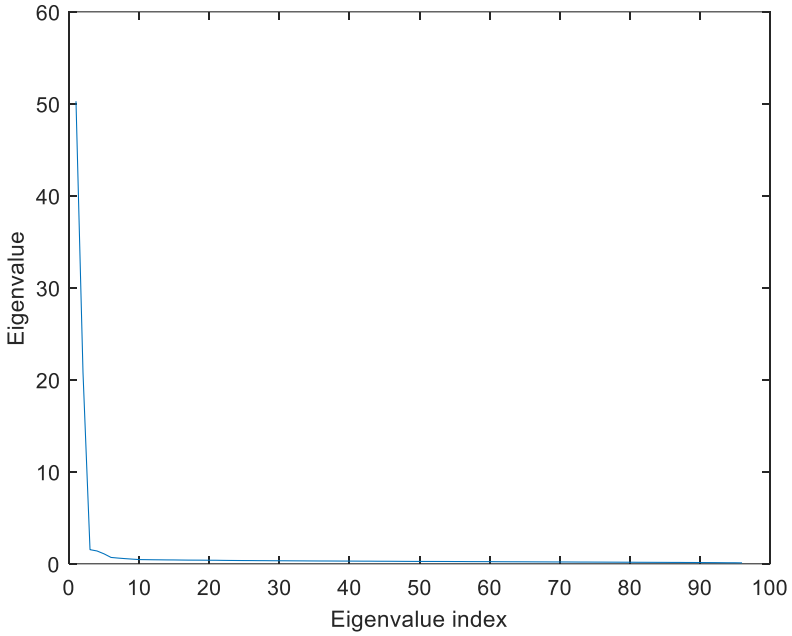

Figure 2.8: Eigenvalues from the PCA from the alternative method.

Once the number of components was determined, we derived the feature importance index by summing the absolute values of the first two components normalized scores. We then sorted the features according importance and proceeded as Step 7 to determine the optimal number of wavelengths. In Figure 2.9, we show the 3D plot of the Silhouette Index according to the tested number of clusters and added features. In the figure, the arrow indicates the highest Silhouette Index, which was with 18 features. One can notice that this number of features is much smaller than our method found (i.e. 41).

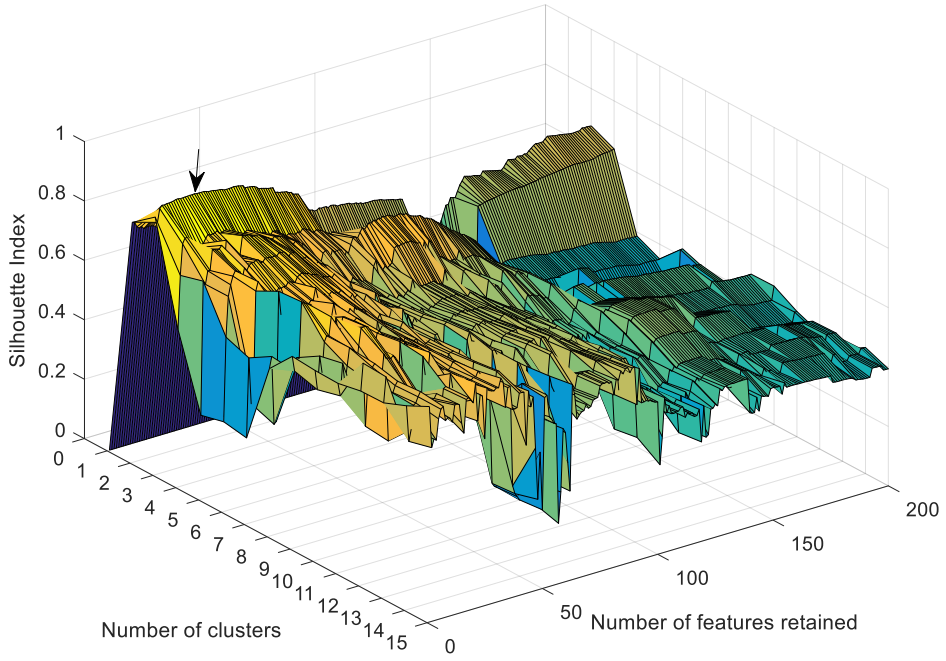

Figure 2.9: Silhouette Index for different number of features and clusters.

To perform a qualitative assessment, we plotted the first 2 PCA components of the selected features according to the comparison method, which is shown in Figure 2.10. When comparing Figure 2.10 to 2.7 (a) and (b), one can notice that this alternative method led to much less clearly separated clusters. Thus, we deem that our method, even selecting more features, presented a much better qualitative performance regarding cluster separability.

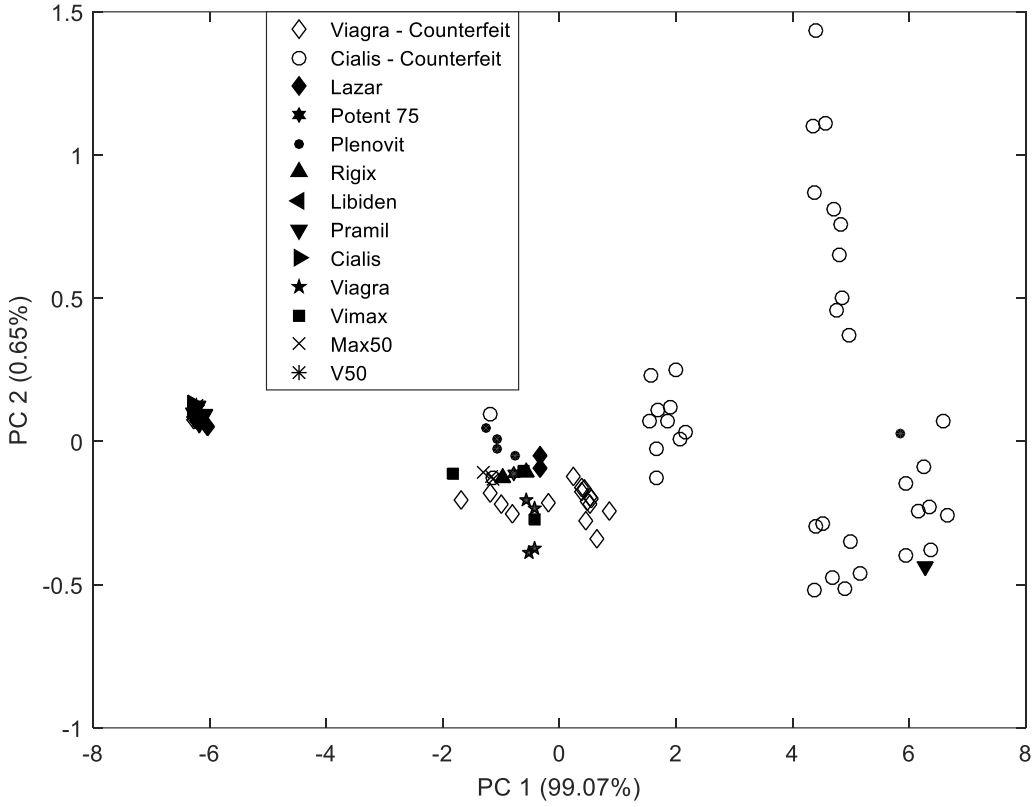

Figure 2.10: Plots of the first two principal components of PCA considering the 18 selected features according to the comparison algorithm.

2.4 Conclusion

The ED-XRF technique is deemed a simple and fast screening technique in forensic analysis to characterize and group authentic samples (commercial formulations of sildenafil citrate and tadalafil) and counterfeit Cialis and Viagra samples. In our study, we included several commercial samples of 11 brands (Viagra®, Cialis®, Lazar®, Libiden®, Maxfil®, Plenovit®, Potent 75®, Rigix®, V-50®, Vimax®, and Pramil®), and counterfeit samples of Cialis and Viagra, from 6 seizures, which were analyzed by XRF.

The XRF data was assessed towards a novel feature selection method aiming at enhancing sample clustering quality by identifying the most relevant energy values. Our method is based on MDS and Procrustes Analysis to derive a feature importance index; such index guides a greedy search approach based on optimization to determine the best number of features (i.e. energy values) to be retained in the model for sample grouping.

From the original 2048 data points in the full spectra, we identified a small number of 41 energy values that led to a better clustering, which was assessed by visual inspection

of the PCA plots. We also performed a comparison with a related work in wavelength selection, where our method also outperformed it. The presented results corroborate the robustness and applicability of our feature selection method for clustering, indicating that the proposed method may be included in analytical protocols aiming to identify counterfeit medicines and cluster medicines of illicit origin.

References

[1] B. Huang, M. Xu, Commentary: Combating sale of counterfeit and falsified medicines online: A losing battle, Front. Pharmacol. 8 (2017) 1–4.

[2] J. Coelho Neto, F.L.C. Lisboa, ATR-FTIR characterization of generic brand-named and counterfeit sildenafil- and tadalafil-based tablets found on the Brazilian market, Sci. Justice. 57 (2017) 283–295.

[3] M. Malet-Martino, U. Holzgrabe, NMR techniques in biomedical and pharmaceutical analysis, J. Pharm. Biomed. Anal. 55 (2011) 1–15.

[4] R.S. Ortiz, K.C. Mariotti, N. V. Schwab, G.P. Sabin, W.F.C. Rocha, E.V.R. de Castro, R.P. Limberger, P. Mayorga, M.I.M.S. Bueno, W. Romão, Fingerprinting of sildenafil citrate and tadalafil tablets in pharmaceutical formulations via X-ray fluorescence (XRF) spectrometry, J. Pharm. Biomed. Anal. 58 (2012) 7–11.

[5] C.M. Bishop, Pattern Recognition and Machine Learning , Springer, 2006.

[6] N. Cebi, M.T. Yilmaz, O. Sagdic, A rapid ATR-FTIR spectroscopic method for detection of sibutramine adulteration in tea and coffee based on hierarchical cluster and principal component analyses, Food Chem. 229 (2017) 517–526.

[7] B. Krakowska, D. Custers, E. Deconinck, M. Daszykowski, Chemometrics and the identification of counterfeit medicines—A review, J. Pharm. Biomed. Anal. 127 (2016) 112–122.

[8] L.S. Lawson, J.D. Rodriguez, Raman Barcode for Counterfeit Drug Product Detection, Anal. Chem. 88 (2016) 4706–4713.

[9] F. Been, Y. Roggo, K. Degardin, P. Esseiva, P. Margot, Profiling of counterfeit medicines by vibrational spectroscopy., Forensic Sci. Int. 211 (2011) 83–100.

[10] W. Romao, M.F. Franco, M.I.M.S. Bueno, M.N. Eberlin, M.-A. de Paoli, Analysing metals in bottle-grade poly(ethylene terephthalate) by X-ray fluorescence spectrometry, J. Appl. Polym. Sci. 117 (2010) 2993–3000.

[11] W. Romão, P.M. Lalli, M.F. Franco, G. Sanvido, N. V Schwab, R. Lanaro, J.L. Costa, B.D. Sabino, M.I.M.S. Bueno, G.F. de Sa, R.J. Daroda, V. de Souza, M.N. Eberlin, Chemical profile of meta-chlorophenylpiperazine (m-CPP) in ecstasy tablets by easy ambient sonic-spray ionization, X-ray fluorescence, ion mobility mass spectrometry and NMR, Anal. Bioanal. Chem. 400 (2011) 3053–3064.

[12] I. Borg, P.J.F. Groenen, P. Mair, Applied multidimensional scaling and unfolding, Springer, 2017.

[13] N. Saeed, H. Nam, M.I.U. Haq, D.B. Muhammad Saqib, A Survey on Multidimensional Scaling, ACM Comput. Surv. 51 (2018) 47:1--47:25.

[14] S.T. Birchfield, A. Subramanya, Microphone array position calibration by basis-point classical multidimensional scaling, IEEE Trans. Speech Audio Process. 13 (2005) 1025–1034.

[15] D.G. Kendall, A survey of the statistical theory of shape, Stat. Sci. (1989) 87–99.

[16] J.C. Gower, Generalized procrustes analysis, Psychometrika. 40 (1975) 33–51.

[17] C.J. Sergeant, E.N. Starkey, K.K. Bartz, M.H. Wilson, F.J. Mueter, A practitioner's guide for exploring water quality patterns using principal components analysis and Procrustes, Environ. Monit. Assess. 188 (2016) 249.

[18] A. Malhi, R.X. Gao, PCA-based feature selection scheme for machine defect classification, IEEE Trans. Instrum. Meas. 53 (2004) 1517–1525.

[19] M.J. Anzanello, F.S. Fogliatto, K. Rossini, Data mining-based method for identifying discriminant attributes in sensory profiling, Food Qual. Prefer. 22 (2011) 139–148.

[20] M.J. Anzanello, F.S. Fogliatto, R.S. Ortiz, R. Limberger, K. Mariotti, Selecting relevant Fourier transform infrared spectroscopy wavenumbers for clustering authentic and counterfeit drug samples, Sci. Justice. 54 (2014) 363–368.

[21] M.J. Anzanello, R.S. Ortiz, R. Limberger, K. Mariotti, PLS-DA wavenumber selection for the categorization of medicine samples based on multiple criteria, Forensic Sci. Int. 242 (2014) 111–116.

[22] M.J. Anzanello, K. Fu, F.F. Fogliatto, M.F. Ferrao, HATR-FTIR wavenumber selection for predicting biodiesel/diesel blends flash point, Chemom. Intell. Lab. Syst. 145 (2015) 1–6.

[23] T.N. Tran, N.L. Afanador, L.M.C. Buydens, L. Blanchet, Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC), Chemom. Intell. Lab. Syst. 138 (2014) 153–160.

[24] W.R. Zwick, W.F. Velicer, Comparison of five rules for determining the number of components to retain., Psychol. Bull. 99 (1986) 432.

[25] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.

[26] S.C. Johnson, Hierarchical clustering schemes, Psychometrika. 32 (1967) 241–254.

[27] R.S. Ortiz, K.D.C. Mariotti, B. Fank, R.P. Limberger, M.J. Anzanello, P. Mayorga, Counterfeit Cialis and Viagra fingerprinting by ATR-FTIR spectroscopy with chemometry: can the same pharmaceutical powder mixture be used to falsify two medicines?, Forensic Sci. Int. 226 (2013) 282–9.

ADDITIONAL MATERIAL

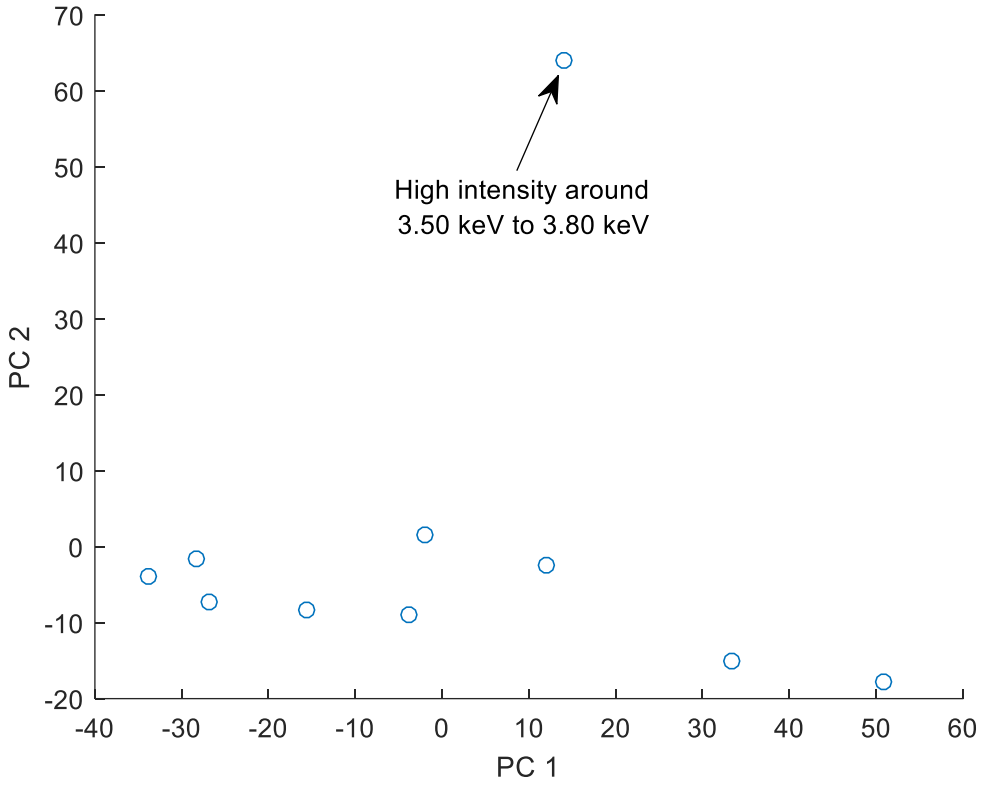

Figure S2.1: PCA plots of the Pramil® samples. The arrow indicates the discrepant sample identified.

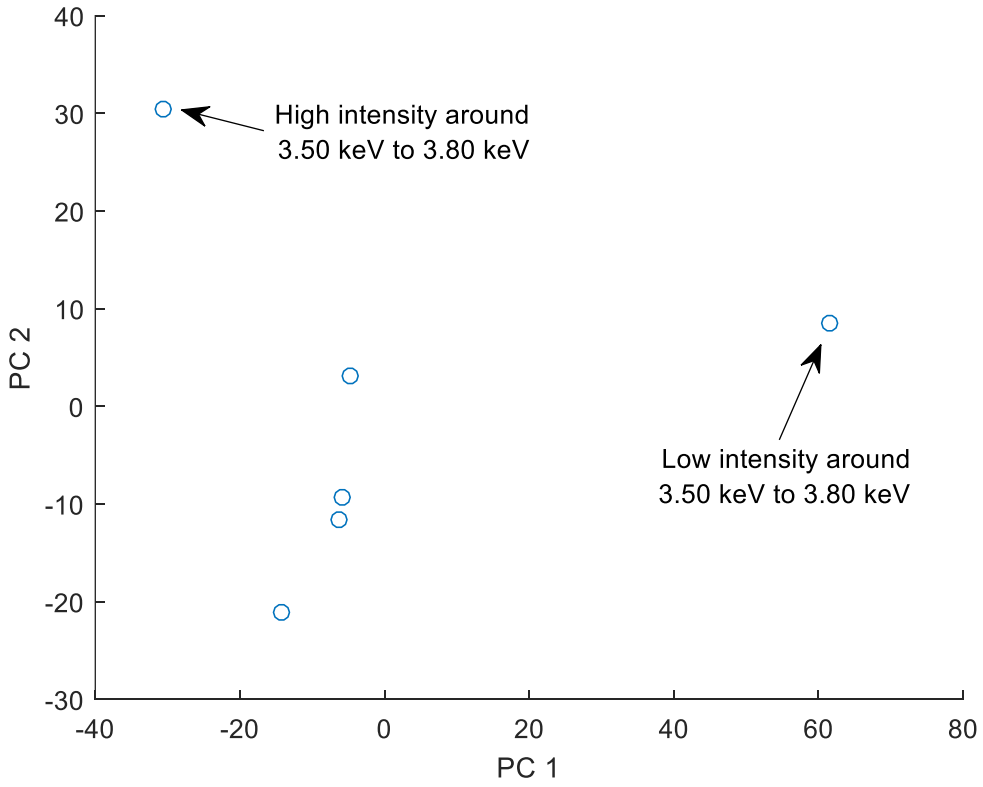

Figure S2.2: PCA plots of the Plenovit® samples. The arrows indicate the discrepant samples identified.

4 ARTIGO 3 - Communication regarding the article "An efficient primary screening COVID-19 by serum Raman spectroscopy"

Published in the Journal of Raman Spectroscopy

Abstract

When performing computational modelling and machine learning experiments, it is imperative to follow a protocol that minimizes bias. In this communication, we share our concerns regarding the article "An efficient primary screening COVID-19 by serum Raman spectroscopy" published in this journal. We consider that the authors may have inadvertently biased their results by not guaranteeing complete independence of test samples from the training data. We corroborate our point by reproducing the experiment with the available data, showing that if full independence of the test set was ensured, the reported results should be lower. We ask the authors to provide more information regarding their article, as well as making available all code used to generate their results. Our experiments are available at https://doi.org/10.6084/m9.figshare.14124356

4.1 Introduction

Due to the COVID-19 pandemic, research and publications have aimed at providing new diagnostic tools for this disease [1,2]. The gold standard RT-PCR may not be extensively available in some countries or not affordable in developing countries. We salute any effort that may help tackle this pandemic, but we as scientists should always be aware of scientific rigor and be sure that bias is reduced during experimentation.

In light of that, we would like to address some methodological aspects regarding the article "An efficient primary screening COVID-19 by serum Raman spectroscopy" [3], published in this journal. The authors proposed using Raman spectra from human blood serum to provide a screening tool regarding COVID-19. They enrolled 177 patients for this study from three different groups: healthy individuals, suspected cases, and COVID-19 positive.

For the analysis, they recorded spectra in the range of 600 - 1800 cm^{-1}. Three experimenters recorded 5 times each sample, resulting in 15 spectra for each subject. The authors mentioned a total of 2355 spectra. After that, 30% of **spectra** (not patients) were set aside for testing, while the remaining 70% for training. The authors performed feature selection using ANOVA to identify the most relevant wavelengths, then used an SVM

classifier to build the discriminant system. They reported accuracy values of 0.87 for COVID-19 versus suspected and 0.91 for COVID-19 versus healthy control.

The previous results on the hold-out set would be impressive. However, we consider that they are likely to be biased due to methodological concerns, which we will now describe, mainly related to violation of training and development/test set independence. Our analysis is based on the published article and their open-access data and code made available at Figshare (https://doi.org/10.6084/m9.figshare.12159924.v1). We would like to draw special attention to this assumption that the codes found in the Figshare are fragments of the proposed approach by the original authors. We tried contacting the original authors, but we received no answer. Thus, we decided to replicate the analysis by following the fragments and reimplementing it, then comparing the results. In addition, the code at (https://doi.org/10.6084/m9.figshare.12159924.v1), was not even running without errors, thus we had to extensive rely on cross-referencing the code fragments and the paper description. We also tried to keep traceability between our implementation and the original released one.

4.2 Concern 1: Supervised wavelength selection with all data or sample leakage

From what was described in the paper (Section 2.4) and the code on Figshare, the ANOVA is carried out on all available data, not only the training set. That means that information from samples on the test set is being "leaked" to the feature selection process, which could by itself introduce a bias towards enhancing classification accuracy.

Hastie et al.[4] explicitly state that first screening the predictors, selecting the most relevant ones, and then cross-validate the classification model gives an unfair advantage. Hastie et al.[4] say that "Leaving samples out *after* the variables have been selected does not correctly mimic the application of the classifier to a completely independent test set, since the predictors 'have already seen' the left out samples".

However, when inspecting the code on Figshare, it does seem that the authors performed the hold-out process before conducting the ANOVA; however, later on in the code, they reshuffle the complete dataset and perform a new hold-out for the classification. Thus, the same case explained above occurs; it is not ensured that the test set samples are not "seen" during the ANOVA-based feature selection process.

4.3 Concern 2: Data for cross-validation may not grouped by patients

The authors stated that a 70/30% cross-validation was performed and repeated 50 times. They also said that "To ensure the independence of the data, the random sampling process guaranteed that the spectra data were used to establish the model and for model test from completely different samples". By checking the published code, we found that the data is being randomly assigned at spectra level, not patient level. Thus, if the initially released code does reflect the article's description, we consider that the way they validated their model goes against their assertion of guaranteeing spectra data independence, and it is not enough to mitigate bias.

The article "Common mistakes in cross-validating classification models" [5], which the authors cited to corroborate their course of action, shows exactly their method described in Sections 2.4 and 3.1 as an example of bias.

To guarantee the independence of the test set, data should be grouped by patients, then patients split into training and test, and finally, their spectra assigned to each partition.

By shuffling all spectra and then performing the cross-validation sampling, they did not ensure that information from the test set was not leaked to the training one via replicates of the same serum, leading to overfitting. Even though due to the heterogeneous nature of human serum and spectra acquisition, a high correlation is expected between different samples from the same subject taken at the same time.

Our point is also corroborated by empirical results from Guo et al. [5], who state that "[...] the dataset should be split at the highest hierarchical level to avoid the over-estimation of classification models". In this study, the highest level would be patient level, not sample or spectra level. Besides, Guo et al. [5] state that by not following a replicate (or this case patient) level cross-validation, the information within the validation dataset is implicitly used during the model construction, violating the independence condition of training and validation.

Considering a hypothetical patient A, with three replicates (A1, A2, and A3), our point is that all three replicates should either lie in training or development, not leaving room for A1 being on training, and (A2, A3) in testing, since they come from the same patient.

In Figure 4.1, we depict what is our view from the published article and the code fragments on Figshare of the employed process. Then, we also show what we consider would be a more appropriate way of performing validation, ensuring that there is no sample leakage

(i.e. one spectrum of patient A in training set, and another spectrum from the same patient on development set, since the authors mentioned that there were replicates from the same serum sample).

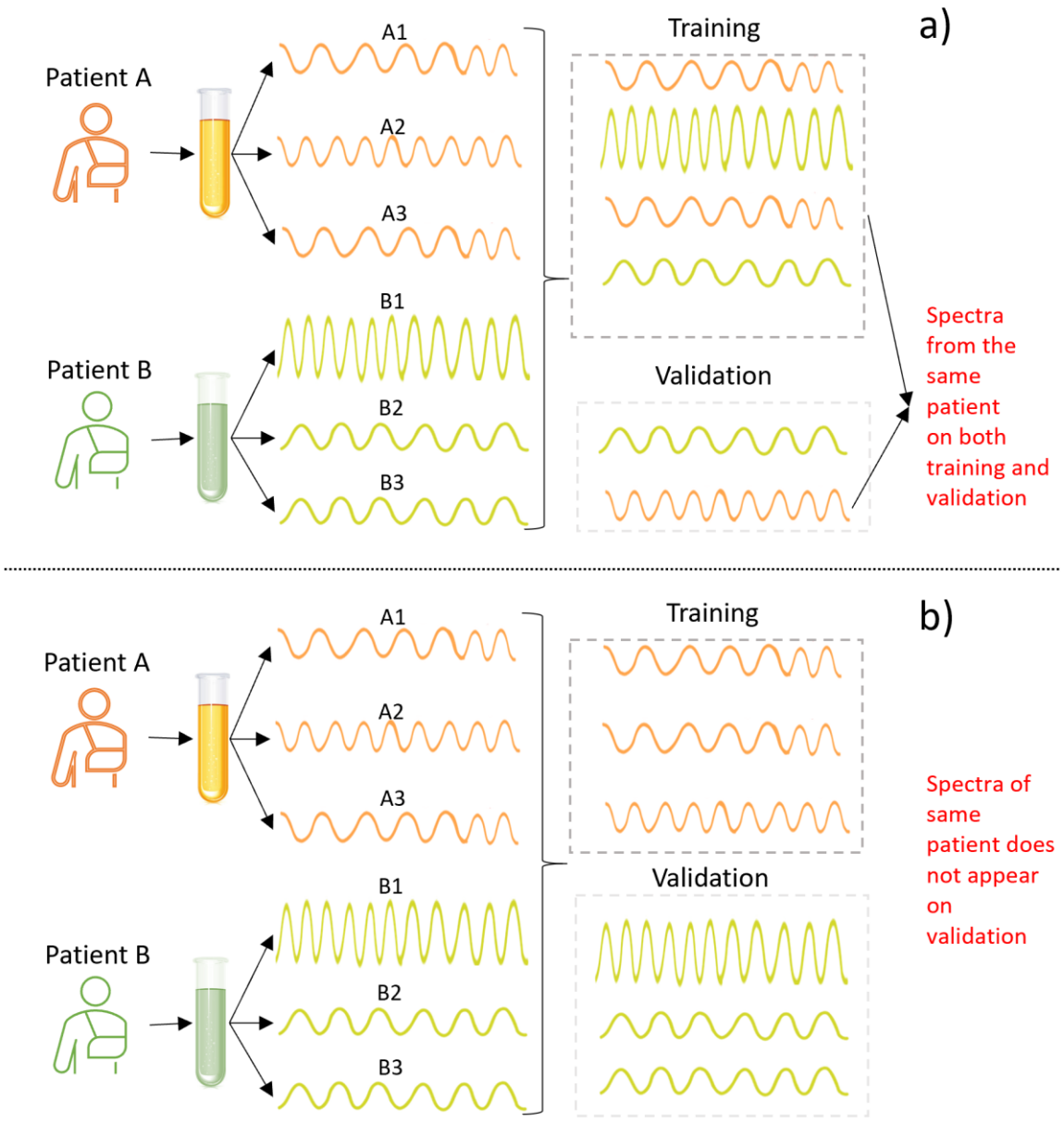

Figure 4.1: Graphical demonstration of how we consider the appropriate way of conducing validation. Please notice that the feature selection step should be included in training, not as a separate step. In (a) we show how we understood that the authors performed evaluation, with leakage during feature selection and possibly during model training, and in (b) how we consider it should be carried out.

4.4 Simulation

We now try to replicate the authors' experiments with their publicly available data and show the impact that data leakage and lack of independence of training/test had. Given that the code provided by the authors that is possibly linked to the publication in question did not run at all (nor completely mirrored what was described in the research paper), we have to re-implement it. We tried to maintain fidelity as much as possible to both the original code and the research paper. In addition, we tried to provide traceability from both code and paper, such that readers can better understand our design choices and to help clarify the divergences and the weak points we try to shed light on this communication.

Although the authors did not specifically assign patient identification for each spectrum on their published data, they provided information on the number of spectra for each patient and which patients had fewer than 3 spectra. Thus, we implicitly generated the patient ID for the highest-level splitting. We corroborated our assumption by experimenting with an LDA classifier aiming at patients' IDs as labels.

Table 4.1 shows the results obtained for accuracy, sensitivity, and specificity, as these were the performance measures used by the original authors. The column "reproduced" is the reproduction of the experiment according to the original article's description. The column "blocked by patients" aims at addressing both concerns 1 and 2 by performing cross-validation at the patient level. That is, first assigning patients to either train or test and then retrieve their spectra. The last column contains the values reported by the authors in their article. We performed a non-parametric Wilcoxon rank-sum test to assess statistical significance between the reproduced and blocked columns over 500 repetitions. All data is available at https://doi.org/10.6084/m9.figshare.14124356.

When comparing the three columns, one can notice that the larger difference in the results are between the values reported on the original paper and the reproduced ones. When considering only accuracy, the difference in COVID versus healthy is around 0.10, which is already a salient deviation. However, when focusing on sensitivity and specificity, we can see that they greatly diverge from the originally reported, in both reproduced experiments and when blocking by patient. When looking only at specificity, which is the ability to rule-out having the disease, the performance for the three comparison groups is strikingly different, going from 0.93 to 0.64, in COVID versus healthy, and from 0.86 to 0.66 in COVID versus suspected.

Table 4.1: Classification performance on reproduced experiment and unbiased cross-validation splitting with experiments blocked at patient level.

Class	Parameter	Reproduced	Blocked by patient	Original Reported
COVID vs. Healthy	Accuracy	0.81 ± 0.12	0.82 ± 0.10	0.91 ± 0.04
	Sensitivity	0.97 ± 0.06	**0.95 ± 0.07**	0.89 ± 0.07
	Specificity	0.64 ± 0.26	0.68 ± 0.21	0.93 ± 0.06
COVID vs. Suspected	Accuracy	0.81 ± 0.12	**0.76 ± 0.12**	0.87 ± 0.05
	Sensitivity	0.97 ± 0.07	0.96 ± 0.08	0.89 ± 0.08
	Specificity	0.66 ± 0.25	**0.56 ± 0.24**	0.86 ± 0.09
Suspected vs. Healthy	Accuracy	0.73 ± 0.11	**0.64 ± 0.09**	0.69 ± 0.05
	Sensitivity	0.88 ± 0.11	**0.82 ± 0.20**	0.70 ± 0.09
	Specificity	0.57 ± 0.26	**0.45 ± 0.27**	0.66 ± 0.09

Note: Values in bold represent statistically significant difference between the reproduced and blocked columns. All p-values were less than 0.01.

Looking at the reproduced results and the experiments blocked by patient, we can see that for COVID versus healthy, only sensitivity was found to be significantly different from the reproduced experiments. This does not repeat when looking at COVID versus suspected, and suspected versus healthy, where all metrics, but sensitivity in COVID versus suspected, were significantly different. Our experiments provide evidence to our claim that the authors' experiments may be biased due to the non-independence of the test set. Furthermore, we can see that specificity values found in our experiments are not in pair with the ones reported by the authors.

As an additional experiment, we also conducted similar analysis, by averaging the spectra at patient level, having the same number of spectra as patients. This was carried out to investigate if there is any additional benefit by averaging the spectra at patient level, rather than sample level. In Table 4.2 we report the results.

Table 4.2: Classification performance on patient average spectra for the replication of the original article and the experiments with blocking by patients.

Class	Parameter	Reproduced	Blocked by patient
COVID vs. Healthy	Accuracy	0.82 ± 0.12	0.83 ± 0.12
	Sensitivity	0.93 ± 0.08	0.94 ± 0.10
	Specificity	0.71 ± 0.26	0.72 ± 0.25
COVID vs. Suspected	Accuracy	0.81 ± 0.12	0.84 ± 0.09
	Sensitivity	0.83 ± 0.22	0.76 ± 0.22
	Specificity	0.80 ± 0.24	0.92 ± 0.08
Suspected vs. Healthy	Accuracy	0.65 ± 0.11	0.66 ± 0.10
	Sensitivity	0.80 ± 0.16	0.80 ± 0.16
	Specificity	0.50 ± 0.28	0.50 ± 0.26

We can see that the results shown in Table 4.2 are very similar to the ones reported in Table 4.1, which means that further averaging the spectra at patient level may not provide additional benefit. This may be related to the nature of SVM, which tends to average support vectors, which may be implicitly already "averaging" spectra during training, if such samples

are support vectors. Thus, even if that was the path followed by the original authors, it would still not have accounted for the discrepancy in the published results.

Overall, we consider that one of the main outcomes of our analyses is that we were unable to completely reproduce the original results published in Guo et al. [5] , although we still found evidence that there's a significant difference in the reported metrics when considering their described feature selection method and by not blocking the experiments by patient. This happens primarily in COVID versus suspected, and Suspected versus healthy group. The reason for that, we hypothesize, is that the signal-to-noise ratio is greater in infected patients, making it easier to both identify the most relevant features, and also perform the final classification.

4.5 Conclusion

In this communication, we aimed at questioning the methodological procedures followed by Yin et al. [3] in their published article. Our main goal is to shed light on the importance of following strict protocols that guarantee test set independence when training machine learning algorithms. Due to their ability to learn complex relationships, it is not hard to overfit a model and bias the final result. The authors did not explicitly define their data processing pipeline, as carried out in related works[6,7], thus leaving room for multiple interpretations. Besides, the early code released on Figshare points towards a different direction from what was described in the article.

To the best of our ability, we tried to reproduce the methods laid out in their original paper, since we had to re-write the whole code as it was not even running. We found results pointing out that our reproduced experiments perform much lower than what was originally presented. When comparing our reproduced experiments with the approach we consider would reduce the bias (i.e. feature selection inside the training data, and blocking the experiments by patients), the difference in results is not that accentuated.

We ask the original authors[3] to provide more information about their methodology, possibly the final code used to generate the data on the paper, and the spectra of all subjects, with their respective identification. We believe that proper scrutiny on research can lead to reliable experiments and beneficial results.

Acknowledgements

Felipe Soares would like to acknowledge AWS Diagnostic Development Initiative (DDI) initiative for providing computational resources.

References

[1] T. Hou, W. Zeng, M. Yang, W. Chen, L. Ren, J. Ai, J. Wu, Y. Liao, X. Gou, Y. Li, X. Wang, H. Su, B. Gu, J. Wang, T. Xu, *PLoS Pathog.* **2020**, *16*, e1008705.
[2] F. Soares, A. Villavicencio, F. S. Fogliatto, M. H. Pitombeira Rigatto, M. José Anzanello, M. A. P. Idiart, M. Stevenson, *bioRxiv*, **2020**.
[3] G. Yin, L. Li, S. Lu, Y. Yin, Y. Su, Y. Zeng, M. Luo, M. Ma, H. Zhou, L. Orlandini, D. Yao, G. Liu, J. Lang, *J. Raman Spectrosc.* , DOI:10.1002/jrs.6080.
[4] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, Springer Science & Business Media, **2009**.
[5] S. Guo, T. Bocklitz, U. Neugebauer, J. Popp, *Analytical Methods*, **2017**, 9, 4410–4417.
[6] F. Soares, K. Becker, M. J. Anzanello, *Artif. Intell. Med.* **2017**, *82*, 1.
[7] S. Feng, D. Lin, J. Lin, B. Li, Z. Huang, G. Chen, W. Zhang, L. Wang, J. Pan, R. Chen, H. Zeng, *Analyst* **2013**, *138*, 3967.

6 CONSIDERAÇÕES FINAIS

Este capítulo apresenta as conclusões da pesquisa desenvolvida nesta tese, destacando os objetivos, métodos e resultados obtidos. Por fim, são indicadas possíveis alternativas para pesquisas futuras.

6.1 Conclusões

Esta tese teve como objetivo principal o desenvolvimento de novas sistemáticas de seleção de variáveis para aplicação em bancos de dados espectrais colineares em tarefas de aprendizado supervisionado e não-supervisionado. Para tanto, o trabalho foi dividido em quatro artigos com intuito de atingir os objetivos específicos propostos. São eles: (*i*) Propor e validar um índice de importância de variáveis com vistas ao agrupamento de amostras de medicamentos; (*ii*) Propor e validar um índice de importância de variáveis que com vistas à regressão; (*iii*) Demonstrar a importância da correta validação de modelos que incluem seleção de variáveis e a sua facilidade de reprodutibilidade; (*iv*) Aperfeiçoar o método mRMR para que possibilite capturar dependências lineares durante o cálculo de redundância de variáveis; e (*v*) Avaliar o desempenho dos métodos propostos quando aplicados em diferentes bancos de dados e comparados a tradicionais métodos de seleção de variáveis da literatura.

O objetivo (*i*) foi atingido no primeiro artigo, o qual apresentou um novo método de seleção de variáveis baseado na construção de um índice de importância de variáveis derivados da aplicação do escalonamento multidimensional e análise de Procrustes. Tal índice orienta um processo iterativo de inserção das variáveis e assim identificar as variáveis mais relevantes para o agrupamento de amostras de medicamentos originais e falsificados. O escalonamento multidimensional reduz o número de dimensões necessárias para representar um conjunto de dados de forma que a distância entre as amostras seja preservada. Uma das principais vantagens da utilização do escalonamento multidimensional é a sua maior resiliência a amostras espúrias.

A principal motivação prática do artigo 1 é auxiliar a identificação de medicamentos falsificados e encontrar os padrões dessas falsificações que permitam associar diferentes lotes apreendidos a uma fonte fraudulenta comum. Neste estudo, foram analisadas 41 amostras comerciais de 11 marcas de medicamentos para disfunção erétil (Viagra®, Cialis®, Lazar®, Libiden®, Maxfil®, Plenovit®, Potent 75®, Rigix®, V-50®, Vimax® and

Pramil®) e 56 amostras falsificadas de Viagra® e Cialis®. Cada amostra é descrita por 2048 variáveis. Como resultado, foi identificado um subconjunto de 41 variáveis que conduziu a uma melhor qualidade de agrupamento, medido pelo SI. As variáveis selecionadas estão relacionadas ao excipiente fosfato de cálcio dibásico, encontrado nas amostras autênticas do Viagra® mas não nas do Cialis®. Além disso, percebe-se que as amostras de Libiden® e Cialis® possuem perfil semelhante nas regiões selecionadas e por isso podem ser agrupadas. O método proposto apresentou melhores resultados quando comparado com o agrupamento das amostras sem a seleção de variáveis e com outro método reportado pela literatura. Até onde se tem conhecimento, este foi o primeiro estudo a utilizar o índice de importância de variáveis baseado no escalonamento multidimensional e análise de Procrustes como ferramenta de seleção de variáveis para o agrupamento de amostras.

Os objetivos (*ii*) e (i*v*) foram alcançados no segundo artigo, que propôs o uso dos pesos da matriz de projeção da técnica *Localized Sliced Inverse Regression* (LSIR) para a elaboração de um índice de importância de variáveis. A regressão inversa é um conjunto de técnicas de redução de dimensionalidade que considera a variável de resposta depende apenas de combinações lineares (ou projeções) dos preditores iniciais. Este artigo teve como motivação a seleção de variáveis informativas para a construção de modelos de predição mais precisos, visto que os bancos de dados espectroscópicos são caracterizados pela elevada dimensionalidade e dados colineares.

No segundo artigo, 12 bancos de dados espectroscópicos e de domínio público foram utilizados para validar o método proposto. O desempenho do método proposto foi comparado com a predição utilizando todas as variáveis originais, bem como com quatro métodos de seleção de variáveis tradicionais na literatura (iPLS, biPLS, siPLS e SPA-PLS). Os resultados obtidos mostram que, ao avaliar a consistência de desempenho entre o conjunto de treino e de teste, o método proposto apresentou os melhores resultados em 5 dos 12 bancos de dados. Com base na literatura pesquisada, este foi o primeiro estudo a utilizar os pesos gerados pelo LSIR para a construção de um índice de importância de variáveis.

O objetivo (*iii*) foi cumprido no terceiro artigo, onde foi demonstrada a importância de garantir a independência do conjunto de treinamento e teste ao utilizar os algoritmos de aprendizado de máquina, diminuindo o viés ao analisar os resultados. Tal artigo teve como motivação indicar que o processo de seleção de variáveis sempre deve ser considerado quando realizado o particionamento de banco de dados, bem como a importância de se realizar tal particionamento no mais alto nível hierárquico de amostras (quando estas forem

adquiridas em replicata, por exemplo). Além disso, o artigo demonstrou a dificuldade de se realizar a replicação de análises que, em teoria, deveriam ter sido descritas de forma não ambígua e detalhada. O artigo demonstra que os resultados obtidos na replicação do estudo, a qual considerou diversos cenários, não equivalem aos reportados no artigo original e revisado por pares.

O objetivo *(iv)* foi atingido no quarto artigo, onde foi proposta uma melhoria no método mRMR para seleção de variáveis com o objetivo de classificação. O método mRMR em sua forma original busca maximizar a relevância das variáveis selecionadas ao mesmo tempo em que reduz a redundância entre elas. Porém, para o cálculo de redundância, são consideradas apenas comparações par-a-par para a métrica de associação (informação mútua, neste caso). Caso uma variável seja uma combinação linear de 2 ou mais variáveis, tal redundância será "diluída" nas comparações par-a-par, o que pode prejudicar a seleção do menor conjunto possível de variáveis. Como aprimoramento do método mRMR, foi proposto que a computação de redundância seja feita em um espaço dimensional reduzido através da análise de componentes principais. Para tanto, a redundância é dada como a informação mútua entre o primeiro componente principal das variáveis já selecionadas e a variável candidata.

O método descrito no artigo 4 foi avaliado em uma tarefa de classificação em um banco de dados de espectroscopia Raman já utilizado no artigo 3. O objetivo consistiu em classificar pacientes com ou sem COVID-19 através da análise espectroscópica do soro de sangue de tais pacientes. Os resultados de base para comparação foram os replicados no artigo 3, pois fornecem uma estimativa onde o viés experimental é reduzido quando comparado com o artigo original onde os dados foram publicados. Como ferramentas de classificação, foram utilizados SVM, regressão logística e naive bayes. Os resultados de base obtiveram uma área abaixo da curva (AUC) de 92% para pacientes COVID-19 versus controles saudáveis, enquanto o mRMR original alcançou 95%, e o método proposto alcançou 96%. Enfatiza-se que ambos os métodos de mRMR produziram uma relação sensibilidade/especificidade mais equilibrada, sendo que os resultados de base alcançaram 96% de sensibilidade e 67% de especificidade, respectivamente. O mRMR original mostrou uma sensibilidade de 88% e especificidade de 94%, enquanto o aperfeiçoamento proposto alcançou 89% e 94%, respectivamente.

Por fim, com base no que foi exposto acima, conclui-se que esta tese cumpriu todos os objetivos específicos propostos e contribuiu para o avanço dos estudos na área de seleção de variáveis com fins de clusterização, classificação e predição de propriedades de amostras.

6.2 Sugestões para trabalhos futuros

Como possíveis extensões da pesquisa apresentada nesta tese, sugerem-se as seguintes ações para pesquisas futuras:

a) Utilizar diferentes métricas de distância para executar o escalonamento multidimensional;
b) Estudar o emprego de *autoencoders* baseado em transformers para redução não-supervisionada de dimensionalidade;
c) Utilizar o método proposto no artigo 2 com propósito de regressão;
d) Derivar heurísticas para o fatiamento dos dados para a regressão inversa que não sejam somente equidistantes;
e) Avaliar o emprego de técnicas como kernel PCA, ICA ou PLS para captura de relações de redundância durante o processo iterativo do mRMR.