UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

TESE DE DOUTORADO

Poder discriminante de Fisher como critério de seleção de componentes principais em problemas de classificação usando PCA-LDA

Valber Elias de Almeida

João Pessoa – PB - Brasil
2022

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

TESE DE DOUTORADO

Poder discriminante de Fisher como critério de seleção de componentes principais em problemas de classificação usando PCA-LDA

Valber Elias de Almeida*

Tese apresentada ao Programa de Pós-Graduação em Química da Universidade Federal da Paraíba como parte dos requisitos para obtenção do título de Doutor em ciências, com área de concentração em Química Analítica/Quimiometria.

Orientador: **Prof. Dr. Mário César Ugulino de Araújo**

* Bolsista: CNPq

João Pessoa – PB - Brasil
JULHO 2022

Catalogação na publicação
Seção de Catalogação e Classificação

A447p Almeida, Valber Elias de.
Poder discriminante de Fisher como critério de seleção de componentes principais em problemas de classificação usando PCA-LDA / Valber Elias de Almeida. - João Pessoa, 2022.
92 f. : il.

Orientação: Mário César Ugulino Araújo.
Tese (Doutorado) - UFPB/CCEN.

1. Química analítica - Classificação. 2. Discriminabilidade - Critério. 3. Redução de dimensionalidade. 4. Padrões - Reconhecimento. 5. Análise Discriminante - Fisher. I. Araújo, Mário César Ugulino. II. Título.

UFPB/BC CDU 543(043)

Elaborado por WALQUELINE DA SILVA ARAUJO - CRB-15/514

Poder discriminante de Fisher como critério de seleção de componentes principais em problemas de classificação usando PCA-LDA.

Tese de Doutorado apresentada pelo aluno Valber Elias de Almeida e aprovada pela banca examinadora em 28 de julho de 2022.

Prof. Dr. Mário César Ugulino de Araújo
Departamento de Química – CCEN/UFPB
Orientador/Presidente

AMANDA CECILIA DA SILVA
Dra. Amanda Cecília da Silva
Pesquisadora – UFPB-PB
Examinadora

Sófacles Figueredo Carreiro Soares
Dr. Sófacles Figueredo Carreiro Soares
CT/UFPB-PB
Examinador

Prof. Dr. Edvan Cirino da Silva
Departamento de Química – UFPB-PB
Examinador

Prof. Dr. Márcio José Coelho de Pontes
Departamento de Química – UFPB-PB
Examinador

Assinaturas da Banca realizadas em modo Webconferência em 28/07/2022, digitalizadas e certificadas pelo Prof. Dr. Mário César Ugulino de Araújo (SIAPE 0334937

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

APS – algoritmo das projeções sucessivas

DISP - Poder Discriminante

GA – Algoritmo Genético (do inglês *Genétic Algorithm)*

LDA – Análise Discriminante Linear (do inglês *Linear Disctiminant Anlálisys)*

NIR – Infravermelho próximo (do ingles *Nearest InfraRed*)

PCA – Análise por Componentes Principais (do inglês *Principal Component Análisys*)

SPA - algoritmo das projeções sucessivas (do inglês *Sucssecive Projections Algorithm*)

SW – *STEPWISE*

PC – Componente Principal (do inglês *Principal Component*)

FS – *Forward Stepwise*

KNN - K- vizinhos mais próximos (do inglês *k- nearest neighbor*)

VIS – Visível

FSLDA - Foley–Sammon LDA

LLD – Discriminação Linear Laplaciana (do inglês *Linear Laplacian Discrimination*)

EFDC – Critério Discriminante de Fisher Explícito (do inglês *Explicit Fisher's Discriminating Criterion*)

ULDA - Análise Discriminante Linear não correlacionada (do inglês *uncorrelated Linear Disctiminant Anlálisys*)

MLR - Regressão Linear Múltipla (do inglês *Multiple Linear Regression*)

MBC - Manutenção Baseada em Condições

PLS-DA - Análise Discriminante por mínimos quadrados parciais (do ingês *Partial Least Squares – Discriminant analisys)*

SENAI - Serviço Nacional de Aprendizagem Industrial

SPXY – Algoritmo de divisão de amostras baseado em X e Y (do Inglês Sample Partition for X and Y)

TCC- Taxa de Classificação Correta

RESUMO

Este trabalho propõe um novo algoritmo que emprega uma adaptação do critério discriminante de Fisher (nomeado aqui como Poder Discriminante, DISP) para escolher componentes principais (obtidas a partir da Análise por Componentes Principais, PCA), que foi usada para construir modelos supervisionados de Análise Discriminante Linear (LDA) para resolução de problemas de classificação multivariada a partir de dados químicos. Este trabalho preenche uma lacuna na química analítica, adaptando a abordagem bem-sucedida originalmente proposta no contexto de reconhecimento facial, utilizando o critério de discriminabilidade de Fisher. O algoritmo PCA-DISP-LDA proposto foi então aplicado a três problemas analíticos diferentes, e um estudo de caso simulado envolvendo: (I) classificação de óleo de soja em relação a data de validade por espectroscopia no infravermelho próximo (NIR), (II) identificação de adulteração em amostras biodiesel/diesel com óleo de soja por espectroscopia VIS-NIR e (III) classificação de amostras de cachaça quanto a adulteração do envelhecimento em barris de madeira. O método proposto demonstrou vantagens em relação a técnicas já estabelecidas na literatura: PCA-LDA convencional, PCA-GA-LDA e PCA-SW-LDA, apresentando resultados em termos de TCC, parcimônia e simplicidade operacional superiores aos métodos citados, para casos em que as PCs com maiores valores de variância Explicada não possuem poder discriminante de Fisher.

Palavras-chaves Discriminabilidade; Redução de dimensionalidade; Classificação; Reconhecimento de Padrões; Análise Discriminante de Fisher

ABSTRACT

This work proposes a new algorithm that employs an adaptation of Fisher's discriminant criterion (named here as Discriminating Power, DISP) to choose principal components (obtained from Principal Component Analysis, PCA), which was used to build supervisedmodels for Linear Discriminant Analisys (LDA) for solving multivariate classification problems from food and fuel chemical data. This work fills a gap in analytical chemistry, adapting the successful approach originally proposed in the context of facial recognition, using Fisher's discriminability criterion. The proposed PCA-DISP-LDA algorithm was then applied to three different analytical problems, and a simulated case study involving: (I) classification of soybean oil in relation to expiration date by near-infrared (NIR) spectroscopy, (II)) identification of adulteration in biodiesel/diesel samples with soybean oil by VIS-NIR spectroscopy and (III) classification of cachaça samples regarding adulteration from aging in wooden barrels. The proposed method showed advantages in relation to techniques already established in the literature: conventional PCA-LDA, PCA-GA-LDA and PCA-SW-LDA, presenting results in terms of TCC, parsimony and operational simplicity superior to the mentioned methods, for cases in which that the PCs with higher values of explained variance do not have Fisher's discriminant power.

Palavras-chaves: discriminability; Dimensionality reduction; Classification; Pattern Recognition; Fisher's Discriminant Analysis.

Capítulo 1

INTRODUÇÃO

1. INTRODUÇÃO

1.1 CARACTERIZAÇÃO GERAL DO PROBLEMA

As análises químicas instrumentais modernos geram uma grande quantidade de respostas analíticas, em que na maioria dos casos possuem informações multicorrelacionadas, dificultando a interpretabilidade dos resultados e aumentando o tempo de computação na aplicação de ferramentas quimiométricas (Li et al, 2020). Para superar essas desvantagens, o uso de estratégias de seleção e/ou redução de variáveis tem sido amplamente divulgado na literatura (Yun et al, 2019). Além disso, devido a otimização do uso das variáveis, os resultados da análise química em estudo, tendem a ser melhores que os obtidos sem a seleção, uma vez que variáveis interferentes e/ou desnecessárias são removidas (Yun et al, 2019; Pasquini et al, 2018).

Esta filosofia foi bem aplicada na Análise Discriminante Linear (do inglês Linear Discriminant Análisys - LDA). Devido a restrições matemáticas relacionadas a quantidade de variáveis de entrada não ser maior que a quantidade de amostras, normalmente LDA é associada a uma técnica de redução de dimensionalidade, que surgiram visando o contorno desta restrição, suprimindo a informação relativa aos interferentes por diversos métodos (Fisher, 1936). As abordagens descritas na literatura para redução de variáveis de entrada nos modelos LDA são divididos em duas estratégias: (I) Seleção de variáveis, nas quais se incluem os métodos stepwise (SW) (Maugis et al, 2011), Algoritmo Genético (do inglês Genetic Algorithm – GA) (Siqueira et al, 2017) e o Algoritmo das Projeções Sucessivas (do inglês Successive Projection Algorithm - SPA) (Fernandes et al, 2017) ou (II) Transformação de dados a partir de métodos como a Análise por Componentes Principais (do inglês Principal Component Analysis - PCA) (Kaznowska et al, 2018; Zhang et al, 2021; Sem 2021).

Em PCA, a dimensionalidade dos dados é reduzida através da decomposição da matriz de dados em combinações lineares das variáveis originais, gerando novas variáveis completamente ortogonais, chamadas componentes principais (PCs) (Chi et al, 2017, Li et al, 2020; Lee e Jemain, 2021). Entretanto, diferentemente de LDA, as combinações lineares das variáveis originais não são obtidas sob a restrição de maximização da discriminação das classes em estudo (ou seja, a separabilidade máxima entre classes), e sim de modo que essas combinações lineares expliquem a maior parcela da variância dos dados. Por esse motivo, o uso de scores de PCs como dados de entrada para LDA compondo a técnica denominada PCA-

LDA, depende de uma escolha correta dos fatores empregados para descrever o problema em avaliação (Chi et al, 2017, Li et al, 2020).

Historicamente, os scores ótimos a serem utilizados no PCA-LDA para fins analíticos foram definidos com base no uso da variância explicada (Fisher et al, 1936; Tominaga, 1999), no entanto, os scores das primeiras PCs, que concentram a maior parte da variância explicada, nem sempre correspondem à melhor solução para o problema de classificação. Assim, esse critério deve ser utilizado com cautela, pois as informações analíticas de interesse nem sempre são descritas pelas PCs com a maior variância explicada. Em alguns casos, o problema de classificação pode ser melhor descrito por scores com baixa variância explicada (chi et al, 2017; Zheng et al, 2005).

Para selecionar apenas os scores mais relevantes para cada problema de classificação, independentemente da variação explicada, alguns trabalhos híbridos foram publicados nos últimos anos. Eles relataram adaptações da PCA-LDA juntamente com algoritmos de seleção de variáveis, como o Algoritmo Genético (PCA-GA- LDA) (Zeng et al, 2016) e o *stepwis*e no modo *forward* (PCA-FS-LDA) (Zheng et al, 2005; Sampaio et al, 2021), para selecionar os scores com maior poder discriminante. Outra forma de utilização dos scores de PCA em problemas de classificação é reordená-los em termos de relevância, em termos da ordem crescente do valor de significância estatística *p* (chi et al, 2017) ou ordem decrescente da discriminabilidade de Fisher (Morais et al, 2019). No primeiro caso, os scores são adicionados ao modelo LDA até que o valor de significância estatística p seja maior que 0,05, enquanto no segundo caso é adotado o critério de máxima precisão de classificação usando o Classificador do Vizinho mais próximo (do inglês k- *nearest neighbor* – KNN).

1.2 OBJETIVOS

1.2.1 Objetivo geral

Neste trabalho são propostas estratégias de seleção de componentes baseadas no uso combinado do critério discriminante de Fisher com as técnicas de classificação multivariada PCA-LDA.

1.2.2 Objetivos específicos

I. Desenvolvimento de novo algoritmo chamado PCA-DISP-LDA que emprega uma adaptação do critério de Fisher (denominado aqui como poder discriminante, DISP) na escolha de componentes principais (obtidos por PCA) que serão utilizados para construir modelos LDA;

II. Aplicação do algoritmo num estudo de caso simulado visando demonstrar de uma forma controlada as vantagens do algoritmo proposto frente aos já existentes;

III. Aplicação da estratégia em problemas de classificação a partir de dados de classificação de óleos em relação ao vencimento, adulteração de misturas biodiesel/diesel com óleo de seja e detecção da adulteração de cachaças quanto ao tempo de envelhecimento.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

2. FUNDAMENTAÇÃO TEÓRICA

2.1 LDA E CRITÉRIO DE FISHER

Nesta seção, matrizes são indicadas por letras maiúsculas em negrito, vetores por letras minúsculas em negrito e escalares por caracteres itálicos. Além disso, a transposição de um vetor ou matriz é indicada com um T sobrescrito.

A técnica de classificação multivariada LDA (Fisher, 1936) foi proposto por Ronald Aylmer Fisher (1890 – 1962), no ano de 1936. Fisher foi um dos mais importantes desenvolvedores da estatística moderna e frequentemente suas técnicas são aplicadas em problemas quimiométricos.

LDA é comumente definido como um método de compressão de dados, baseado na combinação linear de características no espaço amostral definido pelas variáveis, que atua buscando a maximização da discriminação entre classes. Esta definição pode ser entendida avaliando-se por exemplo o LDA calculado pelo algoritmo Foley–Sammon FSLDA (Manfredi et al, 2018; Foley e Sammon Jr, 1975; Hea et al, 2021). Nesse caso, são extraídos da matriz de respostas vetores característicos w, que promovam a maximização o valor do critério de Fisher, de forma que estes vetores possam ser utilizados posteriormente para predição da classe real de cada amostra, de acordo com a função j descrita na equação 1 (Foley e Sammon Jr, 1975; Yang et al, 2002; Hea et al, 2021).

$$j(w) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{s}_{\mathrm{b}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{s}_{\mathrm{w}}\mathbf{w}} \qquad 1$$

De acordo com a equação 1, a combinação linear das características promovida pelo LDA, busca em outras palavras a direção de melhor separação das classes, direção está, representada por wi, que é o i-ésimo autovetor da matriz C = $(Sw)-1Sb$ referente ao maior auto valor, ou seja, corresponde ao vetor de projeção de LDA, que maximiza o critério de Fisher (Foley e Sammon Jr, 1975; Yang et al, 2002; Chen et al, 2014; Hea et al, 2021).

O critério de Fisher, também reportado como separabilidade de classes ou discriminabilidade Di, é definido como a capacidade de um determinado vetor de discriminar amostras de classes distintas e é definido algebricamente de acordo com a equação 2, como a razão de sBi e sWi (Rozza et al, 2012; Caneca et al, 2006; Hea et al, 2021). Vale ressaltar que historicamente sBi sempre foi definido como dispersão entre classes, porém, sabe-se que de

fato não se trata de um cálculo formal de dispersão. O que há na verdade, é a diferença entre o centro das classes, representada pelas médias mij e mi (equação 6), que por uma questão de simplificação será definido aqui como pseudo dispersão entre as classes.

$$D_i = \frac{S_{bi}}{S_{wi}} \qquad 2$$

Onde s_{Wi} é descrita por:

$$s_{Wi} = \sum_{j=1}^{C} S_{ij} \qquad 3$$

Onde s_{ij} é a dispersão de um dado vetor *w* para a classe *j*, calculado como:

$$s_{ij} = \sum_{k \in I_j} [x_i^k - m_{ij}]^2 \qquad 4$$

Em que x_i^k denota o valor de cada k-ésimo objeto do vetor *w*, e m_{ij} é o valor médio do vetor para a classe *j* calculado como:

$$m_{ij} = \frac{1}{n_j} \sum_{k \in I_j} x_i^k \qquad 5$$

A pseudo dispersão entre as classes s_{Bi} é então definida como:

$$S_{Bi} = \sum_{j=1}^{C} n_j [m_{ij} - m_i]^2 \qquad 6$$

Onde m_i é a média dos objetos do vetor e *n* sempre a quantidade de observações.

Com base na equação 1, verifica-se a justificativa da definição inicial de LDA, em que o valor do critério de Fisher aumenta quando o termo sWi presente no denominador da equação diminui e o termo do numerador SBi aumenta, ou seja, quando a dispersão intra classe diminui e a entre classes aumenta o valor da função j é maximizado.

Um inconveniente presente na configuração do critério de Fisher no LDA, é a inversão da matriz de dispersão dentro da classe, na equação C = $(Sw)-1Sb$. Em casos em que a quantidade de variáveis em uma determinada classe é maior que a quantidade de amostras, a inversão do termo $(Sw)-1$ é impossibilitada, pois o resultado dessa operação resulta numa

matriz singular. Este problema é chamado de “small sample problem” ou problema de amostra pequena, em tradução livre (Loog, 2007).

Algumas formulações foram propostas com o objetivo de se contornar esse problema, dentre elas destaca-se o trabalho de Zhao e colaboradores (Caneca et al, 2006), no artigo é proposto uma extensão do LDA convencional denominada Linear Laplacian Discrimination (LLD), a partir de uma alteração na obtenção do critério de Fisher, a fim de se evitar a etapa de inversão de Sw. A principal justificativa dos autores além da elucidação do problema de amostra pequena, é que o LDA, em teoria, não funcionaria adequadamente em problemas em que os dados não tivessem natureza euclidiana. Dessa forma um novo critério de otimização j é exposto na equação 7.

$$j(w) = \frac{|\mathrm{w}(\mathrm{S_b} - \mathrm{S_w})w^{\mathrm{T}}|}{|\mathrm{w}\mathrm{w}^{\mathrm{T}}|} \qquad 7$$

Os autores ainda analisam o desempenho desta métrica em um problema de reconhecimento facial, entretanto, com base nas justificativas apresentadas por Loog, este critério não representa de fato resultados equivalentes ao critério tradicional de Fisher, inclusive argumentando que este não dispõe de uma medida adequada de discriminabilidade, nem aplicável a problemas de reconhecimento de padrões supervisionado.

Uma outra preocupação em relação ao LDA baseado no critério de Fisher, é relacionada a capacidade de generalização dos seus modelos, quando a quantidade de amostras é reduzida (Loog, 2007). Alguns autores relatam que o universo amostral finito caracterizado pelas amostras de treinamento pode ter sua representatividade diminuída, ao passo em que o critério de Fisher é maximizado. De modo que, os vetores característicos ótimos extraídos com base no critério, devem ser aqueles que minimizem a dispersão dentro da classe, dessa forma estes podem ser falhos em descrever o comportamento das amostras, por falta de variabilidade, levando consequentemente a prováveis erros de classificação.

Baseado nisso, Gao e colaboradores (Gao et al, 2012) propuseram um critério discriminante de Fisher melhorado (EFDC), que leva em consideração de forma explicita a variabilidade dentro da classe. Este critério é definido algebricamente de acordo com a equação 8.

$$j(w) = \frac{\mathrm{w}^{\mathrm{T}}[a\mathrm{s_b} + (1-a)\mathrm{s_d}]\mathrm{w}}{\mathrm{w}^{\mathrm{T}}\mathrm{s_w}\mathrm{w}} \qquad 8$$

Onde a é uma constante que controla o equilíbrio entre a capacidade discriminativa e a capacidade de se modelar a variabilidade da classe, e Sd = X(F-D) X^T em que F é uma matriz diagonal que contém o somatório dos pesos D. Segundo os autores, o método LDA baseado neste critério é capaz de modelar satisfatoriamente a discriminação das classes, sem perder a característica da variabilidade dos dados dentro de cada classe.

Na literatura, são reportadas algumas constatações conflitantes quanto ao uso do critério de Fisher. No artigo "O que há de errado com o critério de Fisher?" (Yang et al, 2002) são apresentados alguns resultados contraditórios a respeito do seu uso. Nele, são comparados resultados de uma classificação por FSLDA e Análise Discriminante Linear não correlacionada (ULDA), em que, segundo os autores, apesar dos vetores característicos obtidos por FSLDA apresentarem valor de critério de Fisher maiores que os vetores obtidos por ULDA, o desempenho na classificação de ULDA é superior.

A razão apontada para este resultado é que, no estudo de caso, os vetores obtidos por FSLDA são altamente correlacionados e consequentemente redundantes, diferentemente do que ocorre com o ULDA. Nesse caso, os autores alertam para o fato de que o critério de Fisher não é absoluto, e propõem sua utilização sempre acompanhada da correlação estatística, evitando assim multicorrelação entre os vetores característicos obtidos por LDA, apesar de terem individualmente uma alta discriminabilidade (Foley et al, 1975).

2.2 ANÁLISE POR COMPONENTES PRINCIPAIS ASSOCIADA AO LDA

O método PCA inicialmente reduz a dimensionalidade dos dados, fazendo combinações lineares das variáveis originais, de modo que estas sejam transformadas em variáveis latentes, que descrevam as direções de maior variância explicada (Allegretta et al, 2020). Na **Figura 1** é possível verificar a decomposição das variáveis originais nas componentes principais. Inicialmente (**Figura 1A**) uma primeira componente principal (PC) é projetada na direção de maior variabilidade dos dados, considerando todas as variáveis do plano dimensional (Chi et al, 2017; Li et al, 2020 Zeng et al, 2016; Zheng et al, 2005).

Em seguida uma segunda componente principal é projetada partindo da origem dos dados ortogonalmente à primeira PC. Vale ressaltar que cada nova componente principal gerada é residual em relação a anterior, ou seja, explica uma parte da variabilidade dos dados que não foi explicada pela antecessora, até que toda a informação útil seja explicada pelas PCs (Chi et al, 2017; Li et al, 2020 Zeng et al, 2016; Zheng et al, 2005).

A ideia principal do PCA é representar os dados originais por um produto de duas matrizes: a matriz de scores e a matriz de loadings de modo que o resíduo seja minimizado. Os loadings são definidos como sendo o cosseno do ângulo formado entre cada uma das PCs geradas enquanto que os Scores como demonstrado na **Figura 1B,** são definidos como a projeção das amostras nos eixos de cada componente principal.

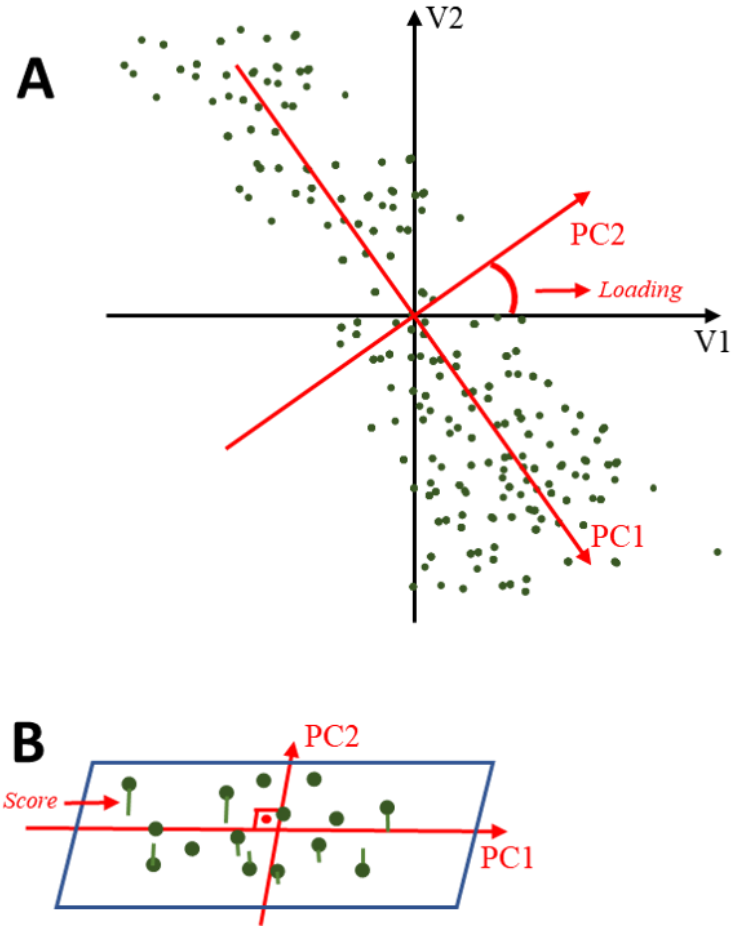

Figura 1 Ilustração do funcionamento da técnica de reconhecimento de padrões de Análise por Componentes Principais (PCA).

Tomando como premissa o fato de que as componentes principais resumem uma porção da informação das variáveis originais, e que como foi demonstrado anteriormente, os scores são representações da projeção das amostras nas componentes principais, os scores podem ser utilizados como dados de entrada de modelos LDA (em vez das variáveis originais), aproveitando o caráter de redução dimensional não multicorrelacionado provocado pela PCA, dando origem à combinação de técnicas para análise de reconhecimento de padrão supervisionada denominada PCA-LDA(Chi et al, 2017; Li et al, 2020 Zeng et al, 2016; Zheng

et al, 2005). Uma descrição matemática mais detalhada da implementação do PCA-LDA é fornecida na seção 4.1.

2.3 ALGORITMO GENÉTICO

O algoritmo genético é uma técnica utilizada para seleção das variáveis que foi implementada a priori, para auxílio na resolução de problemas relacionados a otimização. O algoritmo é classificado como sendo do tipo evolutivo, baseando-se em fenômenos ou observações naturais, para encontrar caminhos otimizados até um dado objetivo. Os principais conceitos em que a técnica se baseia são os de hereditariedade, mutação, recombinação genética e seleção natural de acordo com a teoria da evolução das espécies proposta por Charles Darwin (Leardi e Norgaard, 2004).

O funcionamento do algoritmo se dá a partir da simulação de como uma dada população de indivíduos perpetua suas heranças genéticas, ao longo de um determinado número de gerações. Com isso avalia-se a evolução desta população, iniciada de forma aleatória (característica que atribui o caráter estocástico ao algoritmo genético), e a cada geração é testada a aptidão das populações a resolver um determinado problema, sendo considerada mais apta, as populações que melhor o descreverem. Assim, as populações com maior aptidão são utilizadas como indivíduos de partida para originar a próxima geração (Leardi e Norgaard, 2004).

No contexto da seleção de variáveis, as populações são definidas como as cadeias, constituídas por uma série de indivíduos, que aqui são as próprias variáveis. Estas cadeias de variáveis são então utilizadas para construção de modelos, atribuindo-se maiores valores de aptidão para cada variável (individuo) que melhor descrever o processo. Finalmente as cadeias com maior aptidão geral são utilizadas como ponto de partida para a próxima geração. Nesse caso as variáveis tratadas como indivíduos e são relacionadas com indivíduos de outras cadeias, num processo simulado de reprodução, gerando novas composições de cadeias. Essas novas cadeias são processadas de maneira similar às primeiras até que o número de gerações definido pelo analista seja alcançado. Ao final do processo, a cadeia de melhor aptidão é selecionada (Leardi e Norgaard, 2004).

2.4 STEPWISE

O processo de seleção de variáveis do tipo Stepwise é realizado a partir da checagem da importância de variáveis para predição de uma dada propriedade, por meio de classificação ou calibração multivariada. A inclusão (etapa forward) ou exclusão (etapa backward) de variáveis é definida de acordo com a significância estatística de uma dada medida, associada a cada variável, que atribua melhor desempenho ao modelo (Fisher, 1936; Leardi e Norgaard, 2004; Sampaio et al, 2021).

No processo forward as variáveis são adicionadas a um modelo, enquanto sua adição promover uma melhoria significativa da capacidade preditiva dos modelos. Enquanto que a exclusão das variáveis no processo backward é feito de forma inversa ao passo anterior, excluindo as variáveis menos importantes, de modo que elas sejam excluídas enquanto sua remoção não promover uma redução estatisticamente significativa da capacidade preditiva do modelo (Fisher, 1936; Leardi e Norgaard, 2004; Sampaio et al, 2021).

Capítulo 3

REVISÃO DE LITERATURA

3. REVISÃO DA LITERATURA

Será exposta nesta seção uma discussão a respeito das aplicações e aspectos críticos dos trabalhos reportados na literatura em que o critério de Fisher foi objeto de estudo, demostrando o desenvolvimento e avanços de suas aplicações ao longo dos anos.

No ano de 2008 Groger e colaboradores propuseram a utilização do critério de Fisher para a determinação de perfis químicos que atribuíssem origem geográfica a amostras de heroína e maconha, a partir de dados de cromatografia bidimensional. Como resultado, os autores construíram mapas bidimensionais, contendo os valores de discriminabilidade, para cada amostra analisada em cada tempo de retenção. De modo que nos locais identificados com máxima discriminabilidade, seriam correspondentes aos tempos de retenção do marcador químico de cada tipo de droga, possibilitando a posterior identificação da região de apreensão e origem de cada droga respectivamente (Groger et al, 2008).

Huang e colaboradores utilizaram o critério de Fisher para auxiliar o monitoramento da condição de maquinários industriais no ano de 2010. Esta tarefa é executada a partir da utilização de uma técnica denominada Manutenção Baseada em Condições (MBC), que tem como inconveniente a grande quantidade de variáveis de reposta, o que dificulta a análise direta. Com isso, a partir do critério de Fisher, foi possível identificar quais respostas eram mais discriminantes em termos de conformidade ou não do maquinário em estudo. As variáveis de respostas foram reduzidas com base neste estudo em cerca de 97,7%, aumentando a celeridade e objetividade do monitoramento (Huang et al, 2010).

Cinco anos mais tarde, no ano de 2015 Huang e colaboradores utilizaram o critério de Fisher em problemas de reconhecimento facial, em que dois bancos de dados, sendo um deles composto de 10 fotos de 40 indivíduos em diferentes expressões faciais, perfazendo uma quantidade de 400 imagens, e um outro com 45 imagens de 68 indivíduos em diferentes condições de iluminação e expressões faciais, perfazendo 3060 imagens foram utilizados (Huang et al, 2015).

No trabalho, o critério de Fisher é utilizado para a reordenação de componentes principais, que foram em seguida utilizadas como dados de entrada para classificação por KNN. De modo geral os resultados obtidos para os dois bancos de dados não apresentam um ganho substancial na taxa de classificação, sendo para o primeiro banco de dados um ganho máximo de 1,75%, e para o segundo 2,42%. A principal vantagem da utilização do critério de Fisher no método, é que as PCs utilizadas são as que melhor discriminam as classes, e as PCs não

discriminantes são descartadas, minimizando a possibilidade de sobre ajuste e melhorando a parcimônia (Huang et al, 2015).

Ainda em 2015 Fang e colaboradores propuseram utilizar o critério de Fisher para a redução de dimensionalidade de dados de series temporais, utilizadas na medicina para verificar o avanço de determinadas doenças, fazendo-se uso de técnicas de classificação. As series temporais nesses casos possuem uma alta quantidade de informações, aumentando a probabilidade de conter dados mutuamente redundantes e não informativos, podendo comprometer a eficiência da modelagem, por alta carga computacional e sobre ajuste (Fang et al, 2015).

O método proposto foi testado em dois bancos de dados, sendo o primeiro referente a eletroencefalogramas de um estudo para avaliar a predisposição genética ao alcoolismo, contendo 64 medidas de elétrodos fixados no escalpo. O conjunto de dados foi composto por 100 indivíduos em cada classe (alcoólatras e não- alcoólatras). O segundo banco foi formado por dados do acompanhamento de indivíduos médio-tardios com câncer de pulmão com células não pequenas. Os dados coletados em fichas de históricos médicos possuem 68 índices, e os pacientes foram separados para a composição das classes, de acordo com a evolução tumoral. Com base nos resultados apresentados é possível concluir que o critério de Fisher foi capaz de reduzir a dimensionalidade dos dados de forma otimizada, aumentando a capacidade discriminativa do modelo, em cerca de 93% e 82,4% de acerto para cada um dos bancos de dados respectivamente (Fang et al, 2015).

Já no ano de 2018, Mahdianpari e colaboradores, desenvolveram um método baseado em análise discriminante de Fisher para a classificação das áreas úmidas de três locais, nas províncias de Newfoundland e Labrador no Canadá, a partir de imagens de satélite polSAR. As zonas úmidas ou áreas úmidas são ecossistemas de fronteira entre áreas terrestres e aquáticas e exercem papel importantíssimo na sustentabilidade ambiental de modo que, em escala microscópica fornecem alimentos, armazenam carbono, controle de enchentes, purificação de água, facilitando o escoamento de cursos hídricos, e é o habitat para uma serie de espécies vegetais e animais.

Os resultados obtidos demonstram a eficiência da utilização da matriz de coerência proposta, que foi estabelecida por meio da associação do critério de Fisher e a interpretação física das imagens de satélite. O método proposto apresentou melhores resultados que todas as demais técnicas de classificação estudadas. Segundo os autores, os mapas definidos no estudo, fornecem informações importantes para a preservação destas zonas, facilitando seu monitoramento e fiscalização (Mahdianpari et al, 2018).

No mesmo ano (2018) Kaur e colaboradores utilizaram o critério de Fisher no diagnóstico de tumores cerebrais, visando a melhoraria da taxa de classificação em termos do grau de avanço. No trabalho, dados de 50 indivíduos foram analisados, os quais, 27 apresentavam glioma de baixo grau, e 23 glioma de alto grau. A classificação foi feita a partir de dados de ressonância magnética, permitindo que fosse realizada uma inspeção visual de vários metabolitos no cérebro, auxiliando a verificação de eventuais tumores. A aplicação do critério de Fisher no estudo visa a resolução de três pontos principais: 1- a classificação de tumores cerebrais envolve uma quantidade grande de metabolitos difíceis de se calcular; 2- os espectros obtidos a partir da ressonância magnética requerem uma serie de pré-processamentos; 3- alto custo computacional (Kaur et al, 2018).

De acordo com os autores, uma taxa de classificação correta de 93,72% foi alcançada utilizando KNN como classificador. Uma das vantagens obtidas pelo método proposto é a sumarização das informações obtidas a partir dos metabolitos, sem qualquer necessidade de pré-processamento dos dados. Foi observado ainda, um aumento na precisão do diagnóstico em 4,7%, quando comparado com a classificação convencional (Kaur et al, 2018).

Com base na revisão bibliográfica, é inegável o potencial que o critério de Fisher tem na análise estatística de forma generalizada. Mais precisamente na quimiometria, isto pode ser de grande valia, pois apesar de ser uma métrica bastante consolidada, sua utilização, a parte do LDA, em dados químicos ainda não é totalmente difundida, sendo reportadas na literatura poucas aplicações.

Capítulo 4

METODOLOGIA

4. METODOLOGIA

4.1 ALGORITMO PROPOSTO PCA-DISP-LDA

O algoritmo proposto compreende-se a partir das seguintes etapas (Duda et al, 2001):

Etapa 1 - O número máximo de PCs *k* é definido de acordo com a matriz de treinamento $\mathbf{X_{treino}}$ ($m \times n$). *k* é ($m - 1$) para ($m \geq n$) ou ($n - 1$) para ($n \geq m$).

Etapa 2 - A matriz de treinamento $\mathbf{X_{treino}}$ é decomposta em *k* PCs, que são definidos pelo produto entre a matriz de *scores* $\mathbf{t_{treino}}$ e a matriz transposta de *loadings* $\mathbf{L_{treino}}$.

$$\mathbf{X_{treino}} = [\mathbf{t_{treino}}\mathbf{l}_{\mathbf{treino}}^{\mathbf{T}}]_1 + [\mathbf{t_{treino}}\mathbf{l}_{\mathbf{treino}}^{\mathbf{T}}]_2 + \cdots + [\mathbf{t_{treino}}\mathbf{l}_{\mathbf{treino}}^{\mathbf{T}}]_k + \mathbf{E} \quad \text{Equação 9}$$

Etapa 3 - A matriz de *scores* $\mathbf{T_{teste}}$ do conjunto de testes ($\mathbf{X_{teste}}$) é calculada a partir da matriz de *loadings* $\mathbf{L_{treino}}$ obtida na etapa 2:

$$\mathbf{T_{teste}} = \mathbf{X_{teste}}\mathbf{L}_{\mathbf{treino}}^{\mathbf{T}} \quad \text{Equação 10}$$

Etapa 4 - A discriminabilidade D_i do *score* $\mathbf{T_i}$ referente a cada componente principal PC_i é calculada e em seguida, classificada em ordem decrescente de discriminação. D_i é definida na seção 2.1 como (Duda et al, 2001):

$$\boldsymbol{D_i} = \frac{\boldsymbol{s_{Bi}}}{\boldsymbol{s_{Wi}}} \quad \text{Equação 11}$$

Etapa 5 - Os *scores* do conjunto de treinamento $\mathbf{T_{treino}}$ são usadas como dados de entrada para a construção do modelo LDA. O número ideal de *scores* é escolhido levando-se em consideração a menor taxa de erro obtida pela validação cruzada completa *leave-one-out*.

Etapa 6 – O modelo LDA construído a partir de $\mathbf{T_{treino}}$ é finalmente usado para prever a classe das amostras de teste, empregando os *scores* $\mathbf{T_{teste}}$.

4.2 CONJUNTO DE DADOS

4.2.1 Dados simulados

O conjunto de dados simulados foi composto por 90 amostras, que foram divididas igualmente em duas classes. A simulação foi realizada levando em consideração que a variabilidade na região espectral responsável pela discriminação é muito menor do que a variabilidade das outras regiões espectrais não discriminantes. Essas regiões não discriminatórias apresentam variação aleatória e não têm capacidade de separar as duas classes estudadas.

Cada amostra corresponde a uma mistura de sete constituintes, cujos perfis espectrais puros são mostrados na **Figura 2a**. A região associada ao constituinte C1 (em verde) é a única que contém as informações discriminativas. Por outro lado, as regiões nas quais os outros seis constituintes interagem não possuem informações relevantes para fins de classificação. As concentrações dos constituintes C3, C4 e C6 foram simuladas utilizando um planejamento fatorial 23 em triplicata e treze repetições no ponto central, enquanto a concentração dos constituintes C1, C2, C5 e C7 variou aleatoriamente em torno da média com um desvio padrão característico para cada classe.

A região discriminativa foi definida com a restrição de que a dispersão do constituinte C1 para a classe 1 não se sobrepõe à dispersão de sua concentração na classe 2. Além disso, nessa região, um pequeno desvio padrão foi atribuído ao constituinte C1 para que a variabilidade nessa região seja mantida menor do que em todas as outras. Nas demais regiões espectrais, as concentrações dos constituintes C2 a C7 se sobrepõem e apresentam um desvio padrão muito maior do que o atribuído ao constituinte C1, o que consequentemente aumenta a variabilidade entre as amostras nessas regiões. Isso pode ser visualizado na **Figura 2b**, que apresenta os espectros de todas as 90 amostras e os respectivos desvios padrões para cada variável.

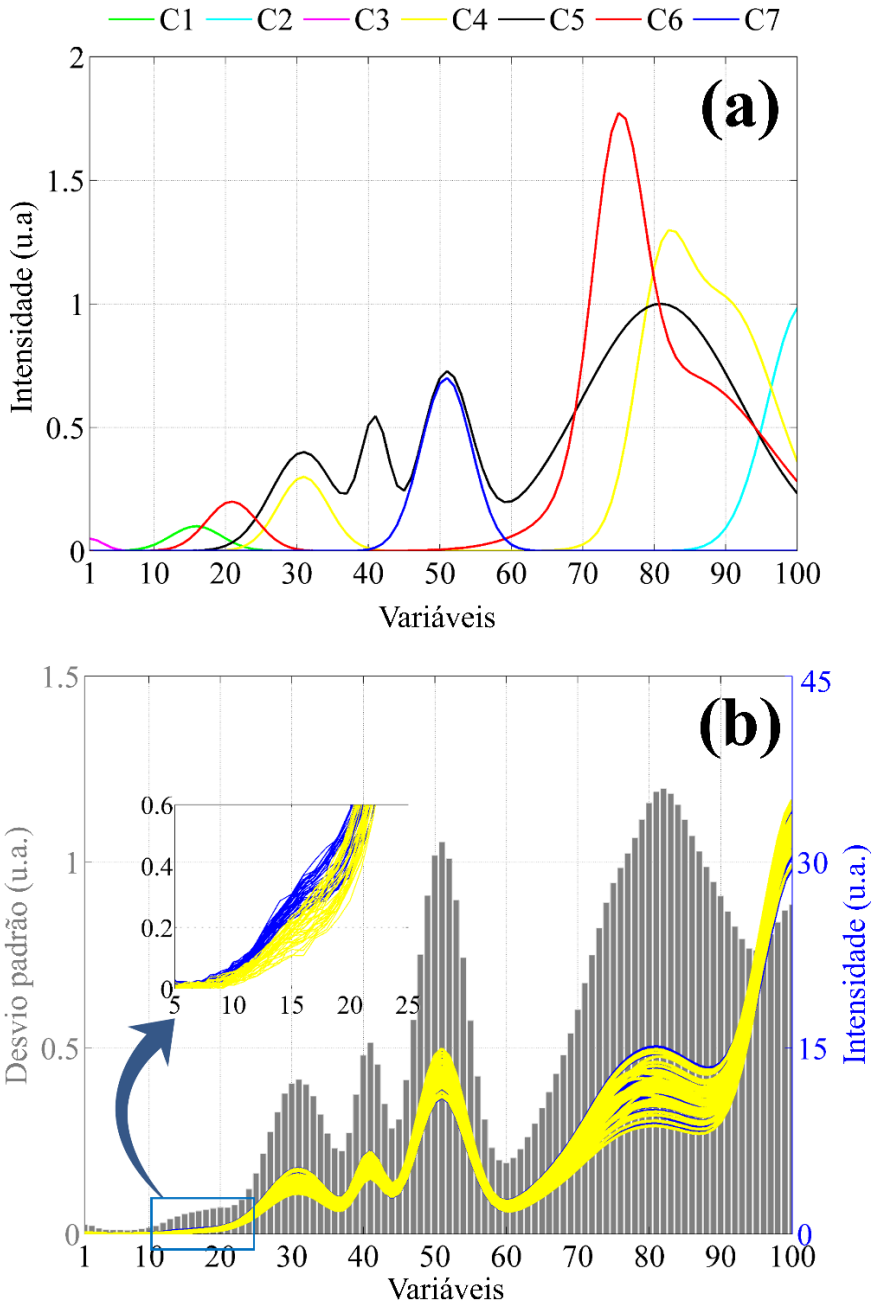

Figura 2 Espectros simulados (**a**) dos sete constituintes puros (**b**) utilizados para gerar as 90 amostras de duas classes diferentes (em azul e amarelo, respectivamente). Os desvios padrões para cada variável são mostrados em barras cinza.

É importante notar que a região discriminativa (entre as variáveis 10 e 20) possui valores de desvio padrão muito menor do que se observa em todo o restante do espectro simulado como destacado no zoom da **Figura 2b**. Isto foi estabelecido de forma proposital, visando a simulação de um sistema, em que a informação discriminante esteja em termos de variância explicada "escondida" na variabilidade geral dos dados, o que provocaria numa eventual análise PCA que

a maior porção da variância explicada seria atribuída a PCs que não possua nenhuma capacidade discriminativa.

De igual maneira, as PCs que sejam fortemente influenciadas pelas variáveis mais discriminativas serão PCs com baixa variância explicada, sendo assim, um cenário perfeito para a demonstração do poder da estratégia que está sendo proposta neste trabalho, uma vez que as PCs serão ordenadas em termos de discriminabilidade e não em termos de variância explicada. Desta forma, espera-se que os modelos sejam mais otimizados usando as componentes principais originais, sem a necessidade de uma rotação nas componentes, como é feito no PLS-DA por exemplo.

4.2.2 Classificação de óleos de soja em relação à data de validade por espectroscopia NIR

Esse conjunto de dados foi composto por 50 amostras de óleos de soja designados para alimentação humana, sendo 30 amostras com medidas coletadas após a data de validade e 20 amostras com medidas espectrais realizadas antes do prazo de validade ter expirado. A fim de se comprovar o estado de oxidação e consequentemente a deterioração do óleo para consumo humano, todas as amostras foram submetidas a ensaios de índice de peróxido, comprovando que cada amostra pertenceria a cada uma das classes previamente indicadas.

Os espectros NIR foram pré-processados visando a correção de flutuações espectrais na linha de base do tipo offset. Todas as medidas referentes às amostras foram realizadas em triplicata. As medidas espectrais na região do infravermelho próximo foram feitas utilizando um espectrofotômetro Perkin Elmer 750 Lambda, equipado com célula de quartzo de caminho óptico de 1 cm, fonte de tungstênio e tubo fotomultiplicador R928 e sistemas de detecção de PbS refrigerados a Peltier, na faixa de 1100 a 1600 nm com resolução espectral de 1nm (**Figura 3**). Para mais detalhes em relação ao procedimento experimental do banco de dados consultar a referência (Costa et al, 2016).

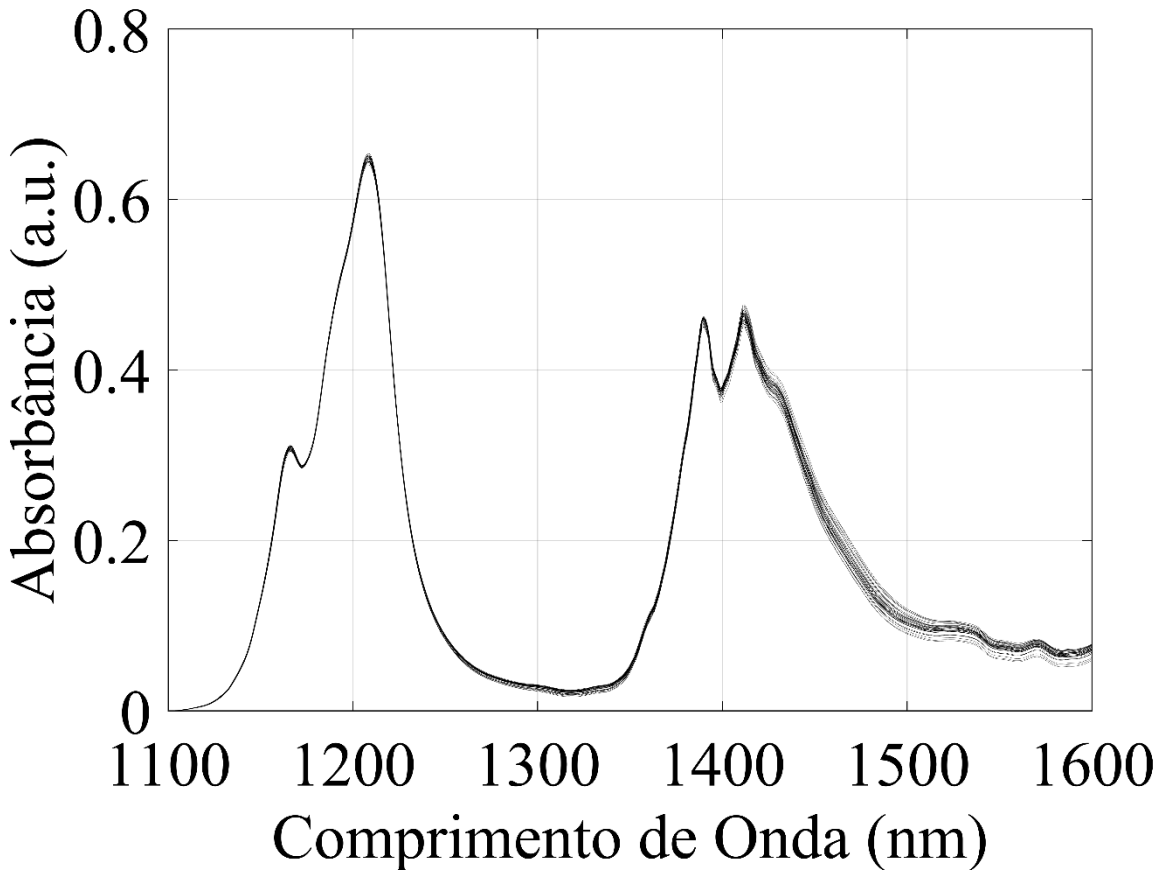

Figura 3 Espectros de absorção molecular obtidos a partir de espectroscopia NIR para amostras de óleos de soja em relação à vida útil: não expirados e expirados.

4.2.3 Identificação da adulteração de misturas de biodiesel/diesel com óleo de soja por espectroscopia Vis-NIR

Este conjunto de dados contendo 90 amostras de misturas de biodiesel/diesel B5 foi utilizado em um trabalho anterior (Fernandes et al, 2014), sendo 31 amostras autênticas (não adulteradas) e 59 amostras adulteradas com óleo de soja na faixa de 0,5 a 2,5% (vv^{-1}) Essas proporções correspondem de 10 a 50% da substituição do biodiesel pelo óleo de soja na mistura B5 (**Figura 4)**. Na referência (Fernandes et al, 2014), apenas a faixa espectral visível (430-650 nm) foi apresentada.

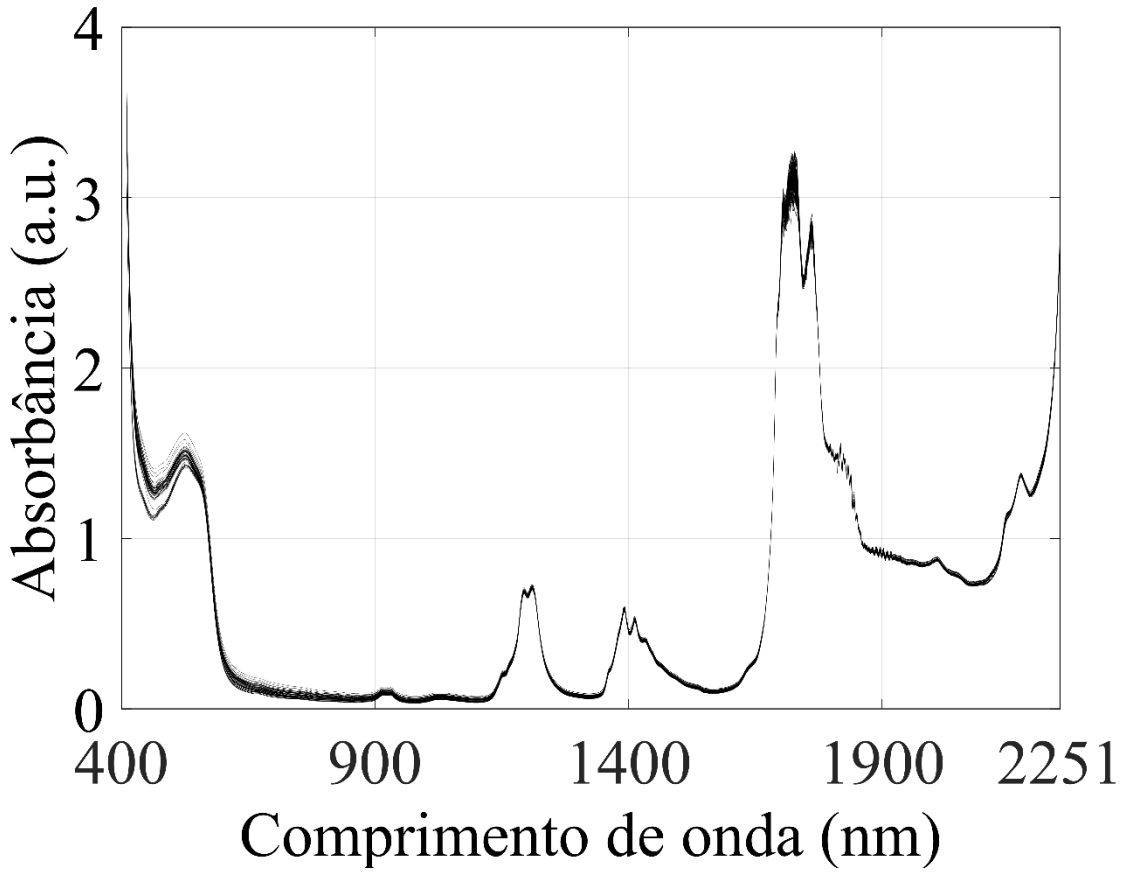

Figura 4 Espectros de absorção molecular obtidos a partir de espectroscopia Vis-NIR para misturas de biodiesel/diesel B5: não adulteradas e adulteradas com óleo de soja na faixa de 0,5 a 2,5% (v / v).

Aqui também foi incluída a faixa espectral do NIR, mantendo regiões saturadas e ruidosas (1650 a 2000 nm), além de não executar nenhum pré-processamento no conjunto de dados, visando avaliar a capacidade do método proposto de evitar componentes principais com alta influência das variáveis correspondentes a estas regiões de saturação e baixa relação sinal/ruido.

As medidas espectrais na região do infravermelho próximo e visível foram feitas utilizando um espectrofotômetro Perkin Elmer 750 Lambda, equipado com célula de quartzo de caminho óptico de 1 cm, fonte de tungstênio e tubo fotomultiplicador R928 e sistemas de detecção de PbS refrigerados a Peltier, na faixa de 410 a 2251nm com resolução espectral de 1nm.

4.2.4 Identificação da adulteração de cachaças envelhecidas por espectroscopia Vis-NIR

Este conjunto de dados foi composto por 91 amostras de cachaça, sendo 61 amostras envelhecida em barris de carvalho (Quercus spp.), amburana (Amburana cearensis) e freijó (Cordia goeldiana), durante o perído de tempo estabelecido em legislação para serem classificadas como chachaça envelhecida e 30 amostras de cachaça recém destiladas (denominadas como cachaça branca) que foram em seguida adicionadas de extratos de madeiras carvalho e freijó, visando simular a aparência física de cachaças autênticas envelhecidas nos respectivos barris de madeira (**Figura 5**).

O teor alcoólico de todas as amostras foi padronizado para 48º utilizando água destilada, de acordo com o que é feito para as amostras autenticamente envelhecidas, visando reduzir a interferência do teor alcoólico nas medidas espectrais no infravermelho próximo.

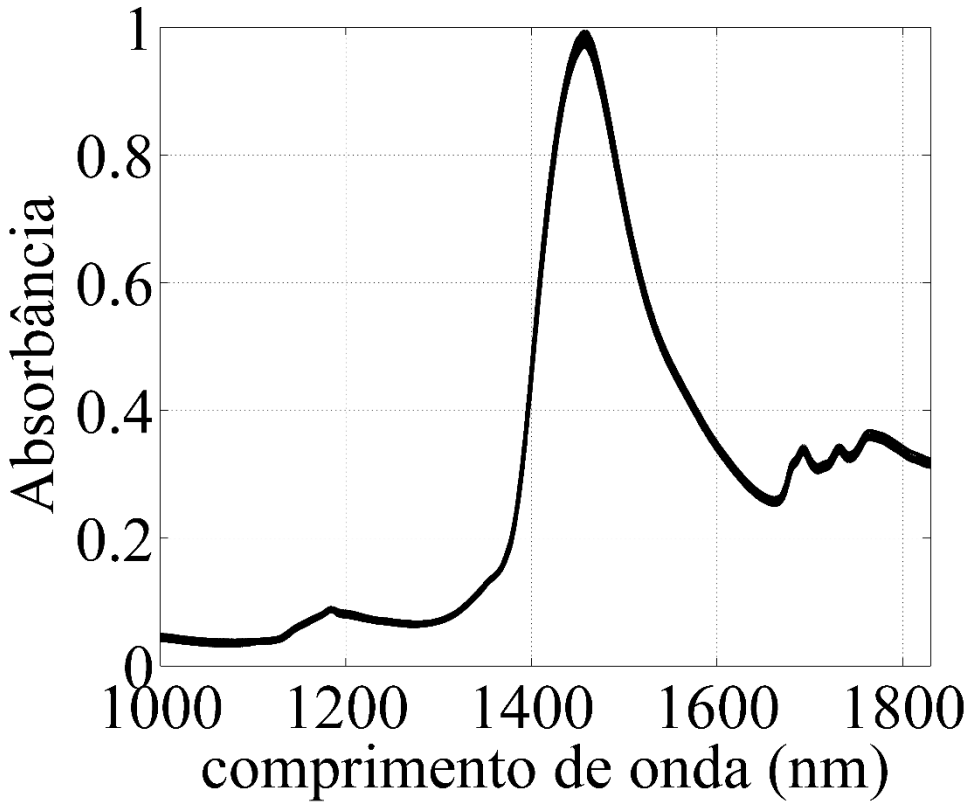

Figura 5 Espectros de absorção molecular obtidos a partir de espectroscopia NIR na região espectral de 1000 a 1800nm para misturas de cachaça envelhecida: não adulteradas e adulteradas com extrato de madeira.

As amostras de cachaça envelhecida e não envelhecida, bem como os extratos de madeira, foram cedidas por destilarias de cana-de-açúcar, localizadas na região produtora da cidade de Areia, Paraíba, Brasil, em um estudo realizado em parceria com o Instituto SENAI de Tecnologia localizado no Distrito Industrial, na cidade de João Pessoa - PB. As amostras foram então armazenadas em frascos âmbar e medidas por espectroscopia NIR (Almeida et al, 2021).

As medições NIR foram realizadas na faixa de 1000 a 1800 nm com resolução de 1 nm usando um espectrofotômetro Perkin Elmer 750 Lambda, equipado com célula de quartzo de caminho óptico de 1 cm, fonte de tungstênio e tubo fotomultiplicador R928 e sistemas de detecção de PbS refrigerados a Peltier.

4.3 PROCEDIMENTO QUIMIOMÉTRICO

Todos os procedimentos quimiométricos foram realizados no Matlab® 2011b. Os conjuntos de amostras foram particionados em conjuntos de treinamento e teste usando o algoritmo SPXY (Galvao et al, 2005), sendo respectivamente 70/20 para os dados simulados, 35/15 para o caso de estudo envolvendo classificação de óleo de soja, 70/20 para o exemplo envolvendo adulteração de biodiesel / misturas a diesel e 76/15 para o conjunto de identificação da adulteração de envelhecimento de cachaça por espectroscopia NIR.

O conjunto de treinamento foi submetido à validação cruzada completa do tipo leave-on-out para escolher o número ideal de componentes principais, adotando como critério a maximização da taxa de classificação correta (TCC). Nos conjuntos de treinamento e teste, mantendo a interpretabilidade dos loadings das componentes principais utilizados na construção dos modelos de LDA.

Para fins de comparação, os resultados do algoritmo proposto PCA-DISP-LDA foram avaliados em relação ao desempenho dos algoritmos PCA-LDA, PCA-GA-LDA e PCA-SW-LDA convencionais. Esses dois últimos modelos foram calculados considerando os seguintes parâmetros de entrada: uma população de 100 indivíduos ao longo de 10 gerações para PCA-GA-LDA e um limiar de correlação de 0,05 para PCA- SW-LDA, para os três primeiros bancos de dados e 100 indivíduos ao longo de 50 gerações para PCA-GA-LDA e um limiar de correlação de 0,1 para PCA-SW-LDA para o banco de dados referente a identificação da adulteração do envelhecimento de cachaça.

Todos os modelos LDA obtidos utilizando as componentes principais selecionadas a partir das respectivas técnicas de seleção de variáveis, foram construídas utilizando a versão 5.4 do classification toolbox (Ballabio e Consonni, 2013).

Capítulo 5

RESULTADOS E DISCUSSÃO

5. RESULTADOS E DISCUSSÃO

5.1 DADOS SIMULADOS

Inicialmente, os dados foram submetidos a uma Análise por Componentes Principais, a fim de se verificar a possibilidade da formação de possíveis agrupamentos naturais, assim como o poder de separação das componentes principais entre as amostras.

As **Figuras 6a-f** exibem os scores das seis primeiras componentes principais (PCs) em ordem decrescente da variância explicada. As cinco primeiras PCs não indicaram uma tendência de discriminação entre as duas classes estudadas, mesmo tendo valores mais altos de variância explicada que a sexta PC (**Figura 6f**).

Este comportamento já era esperado, uma vez que na concepção dos dados simulados, um maior desvio padrão foi atribuído a regiões cuja a influência de constituintes não discriminativos era maior, tendo como consequência PCs com alta variância explicada, mas sem nenhuma discriminação ou agrupamento entre as classes.

Em contrapartida, como é demonstrado no gráfico de scores da PC6, há uma clara discriminação entre as amostras de classes distintas, que embora tenha apenas 0,12% da variância explicada, não se sobrepõem nenhuma amostra de uma classe sobre a outra. Para evidenciar essa descoberta, a discriminabilidade de Fisher foi calculada para cada PC e comparada com as variâncias explicadas para cada PC, conforme ilustrado na **Figura 7**.

De fato, a PC6 tem o maior valor de discriminabilidade, enquanto que as outras estão muito próximas de zero. Em outras palavras, quando as informações relevantes têm uma variabilidade muito menor que as informações não classificatórias, como ruídos, ou informações de constituintes que não atribuem nenhum poder discriminativo, as PCs com maior poder discriminante terão uma baixa variância explicada e, consequentemente, podem não ser adicionadas ao modelo, sendo consideradas resíduos de variância e serem eliminadas junto com a parte não modelada.

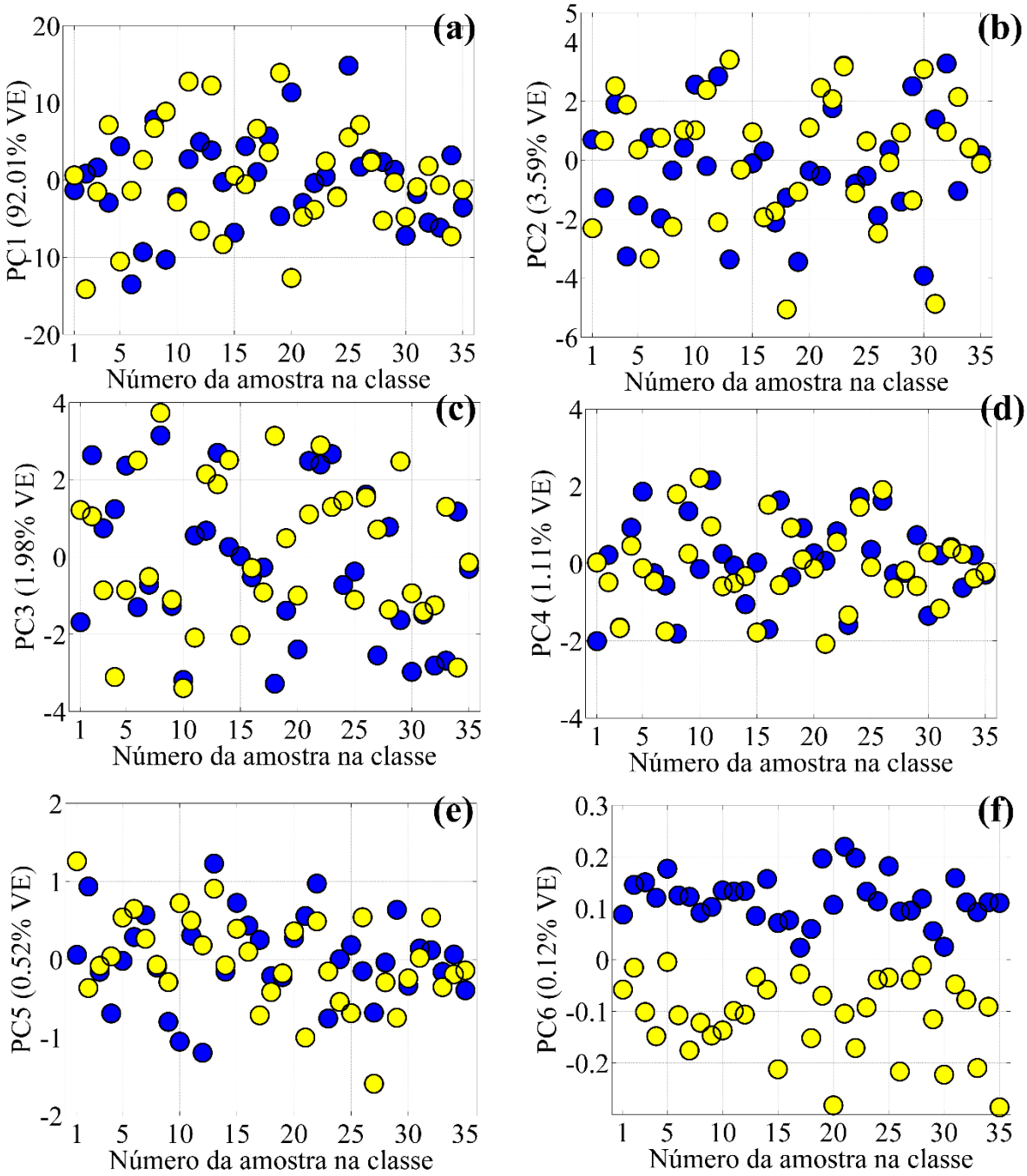

Figura 6 Dados simulados: (a-f) Scores de PCA para as seis primeiras componentes principais (PCs) ordenadas de forma decrescente de variância explicada.

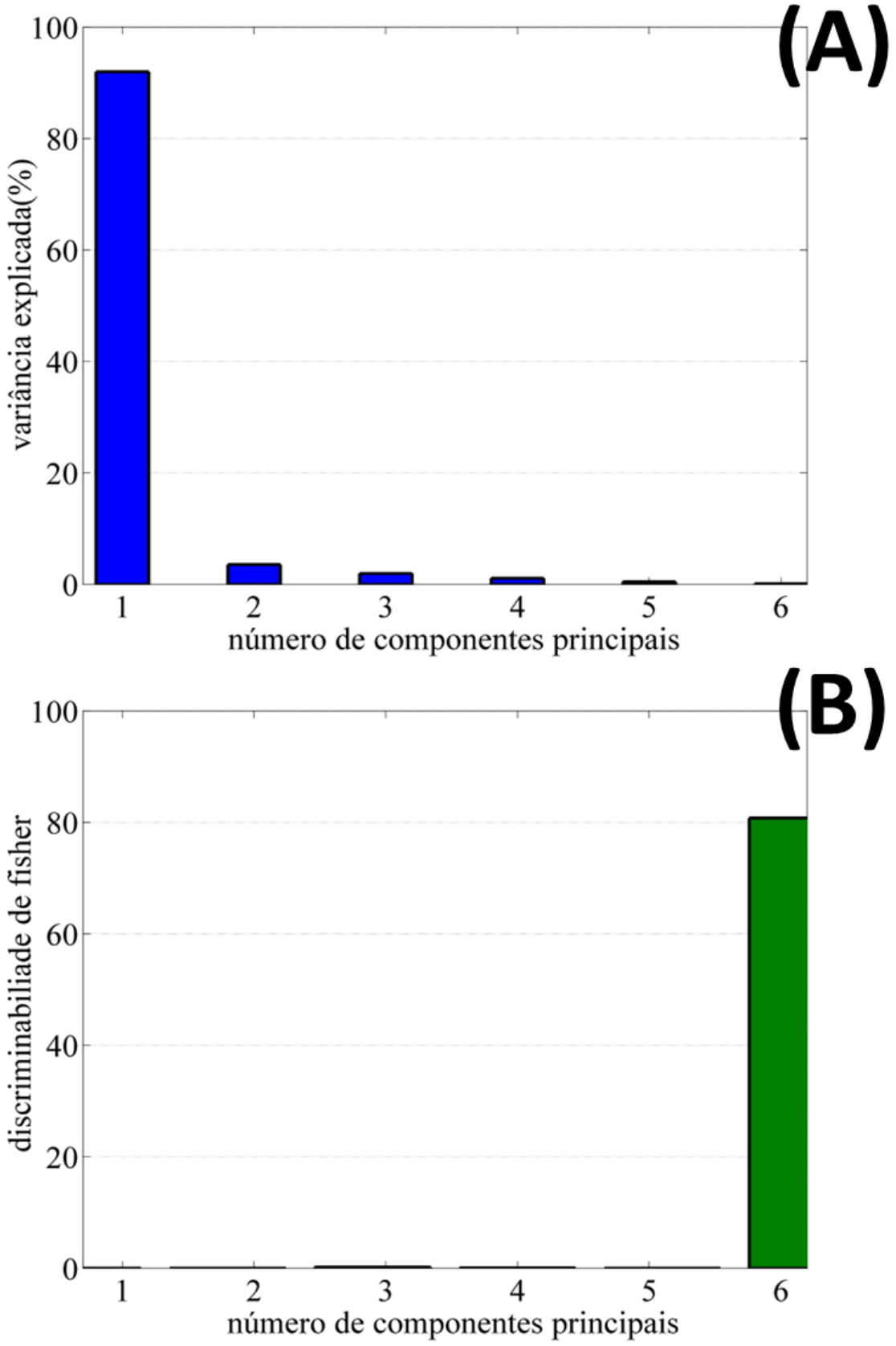

Figura 7 Valores de variância explicada versus discriminabilidade de Fisher para as primeiras seis PCs.

Na **Figura 8** o gráfico de *loadings* para as seis primeiras PCs a partir dos dados simulados são apresentadas. Como pode ser visto, somente na sexta componente principal os *loadings* apresentam informações altamente significativas na faixa espectral (entre as variáveis 10 e 20, neste caso) na qual o componente C1 foi adicionado, contribuindo para a separação completa das amostras em suas respectivas classes.

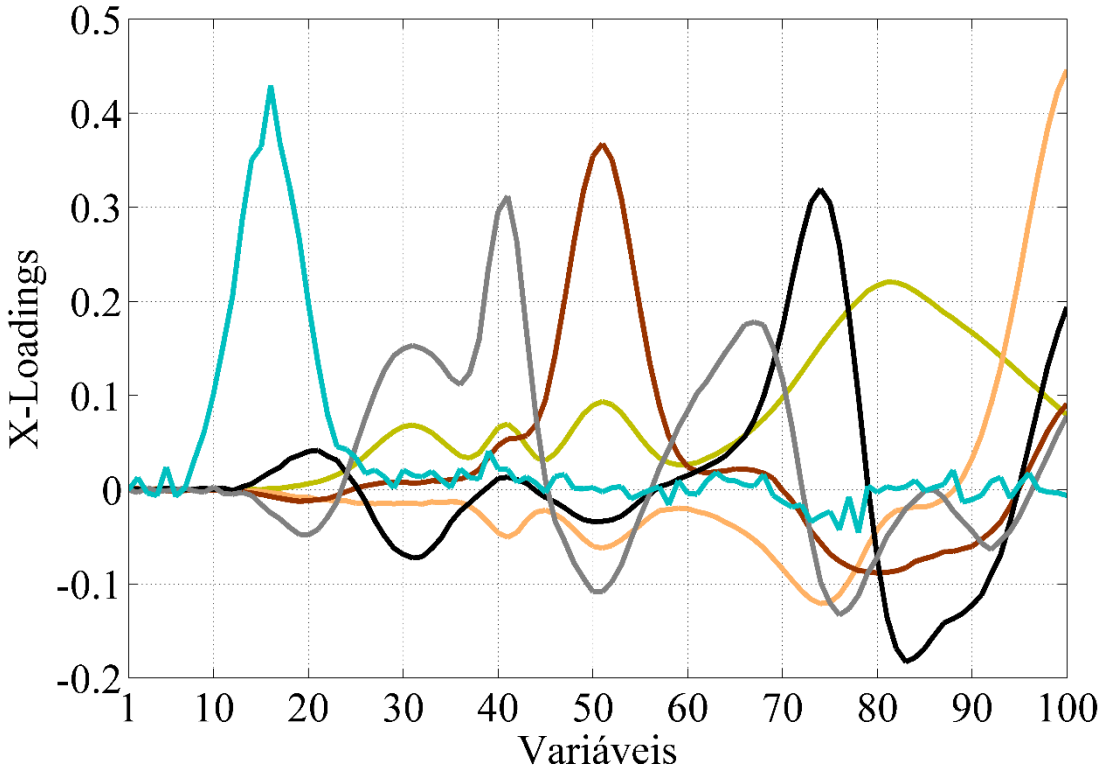

Figura 8 loadings de PCA para as seis primeiras PCs.

Após essa análise exploratória inicial, o modelo PCA-DISP-LDA foi construído usando as amostras de treinamento e sua capacidade de classificação foi avaliada usando apenas as amostras de teste. Na **Tabela 1** são apresentados os resultados da classificação obtidos para PCA-LDA, PCA-GA-LDA, PCA-SW-LDA e PCA-DISP-LDA. Os *scores* e *loadings* de todas as componentes principais utilizadas nos modelos são mostrados no apêndice da seção 8 ao final do documento.

Tabela 1. Resultados da classificação obtidos para PCA-LDA, PCA-GA- LDA, PCA-SW-LDA e PCA-DISP-LDA.

Modelo	PCs (Índice) incluídas no modelo LDA	TCC (%)	
		Treinamento (erros)	Teste (erros)
PCA-LDA	6 (1,2,3,4,5,6)	98.57 (1)	100 (0)
PCA-GA-LDA	7 (2,3,6,8,25,26,33)	98.57 (1)	100 (0)
PCA-SW-LDA	3 (3,6,17)	100 (0)	100 (0)
PCA-DISP-LDA	1 (6)	100 (0)	100 (0)

Como pode ser visto, todos os modelos atingiram 100% da taxa de classificação correta no conjunto de teste. No entanto, como as PCs são ordenadas de maneira decrescente em termos de variância explicada no PCA-LDA convencional, foi necessária a inclusão das cinco primeiras PCs até que a PC6 também seja incluída; as vinte e oito PCs a seguir (ou seja, PC7 a PC34) não foram selecionados para serem incluídos no modelo LDA.

Devido à introdução desses fatores não discriminatórios, os resultados da classificação no conjunto de treinamento foram levemente prejudicados, atingindo um TCC de 98,57% com uma amostra mal classificada (**figura 9**).

Com base nesse resultado, é possível concluir uma premissa defendida na seleção de variáveis, em que a inclusão de variáveis não adequadas no modelo pode prejudicar seu desempenho, uma vez que informações não úteis podem confundir o treinamento do classificador, levando a erros. Além disso, a carga computacional necessária para processar um modelo desnecessariamente pesado e ainda prejudicado reforça a importância do método proposto.

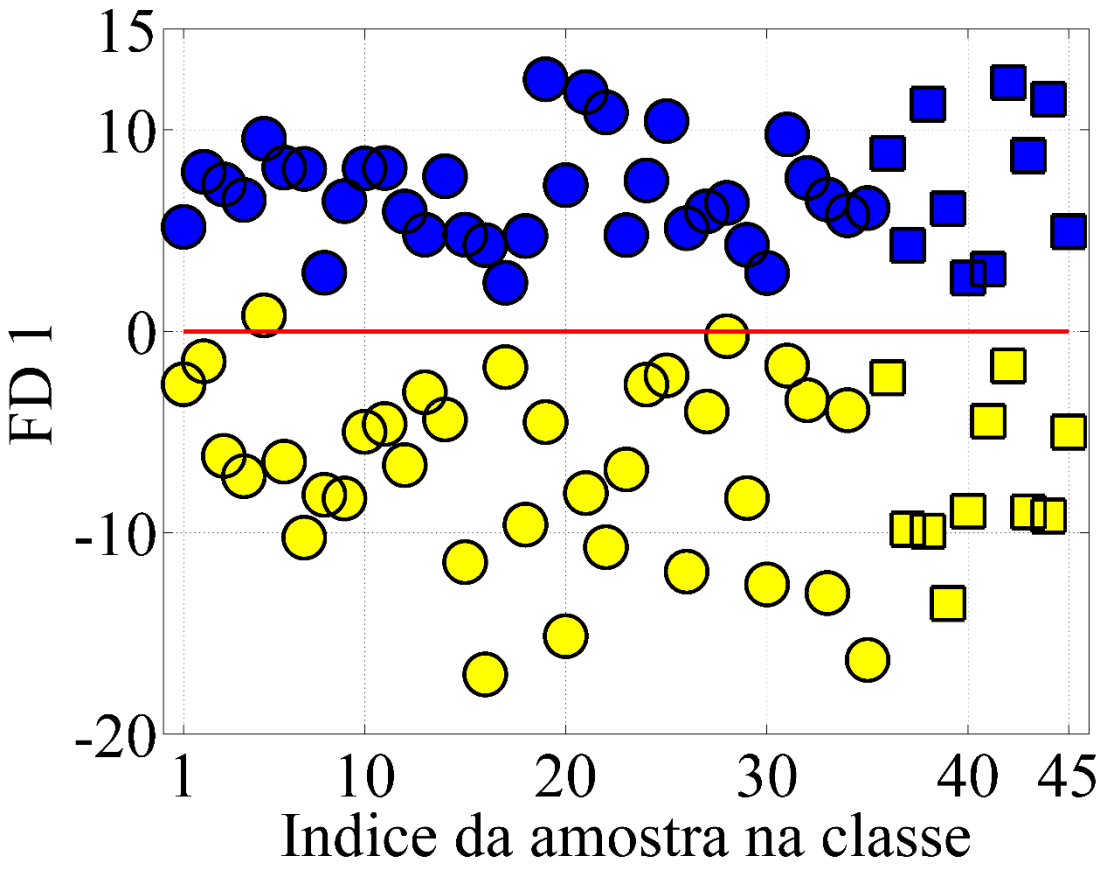

Figura 9 Gráficos de função discriminante de Fisher obtidos para PCA-LDA convencional, A linha vermelha indica o limite entre as classes, enquanto círculos e quadrados representam as amostras nos conjuntos de treinamento e teste, respectivamente, sendo azul para a classe 1 e amarelo para a classe 2.

Levando em conta as técnicas de seleção de variáveis, o método GA selecionou as componentes PC2, PC3, PC6, PC8, PC25, PC26 e PC33. Como sabemos, apenas a componente principal 6 possui informação relevante para a classificação, entretanto o método selecionou mais componentes do que era realmente necessário. Porém, por incluir a PC6 o método ainda apresentou um bom resultado de classificação, embora, de modo semelhante ao que aconteceu com o modelo PCA-LDA, foi prejudicada pela adição de componentes não discriminativas, atingindo TCC de 98,57 e 100% para os conjuntos de treinamento e teste, respectivamente (**Figura 10**) errando uma amostra no conjunto de treinamento.

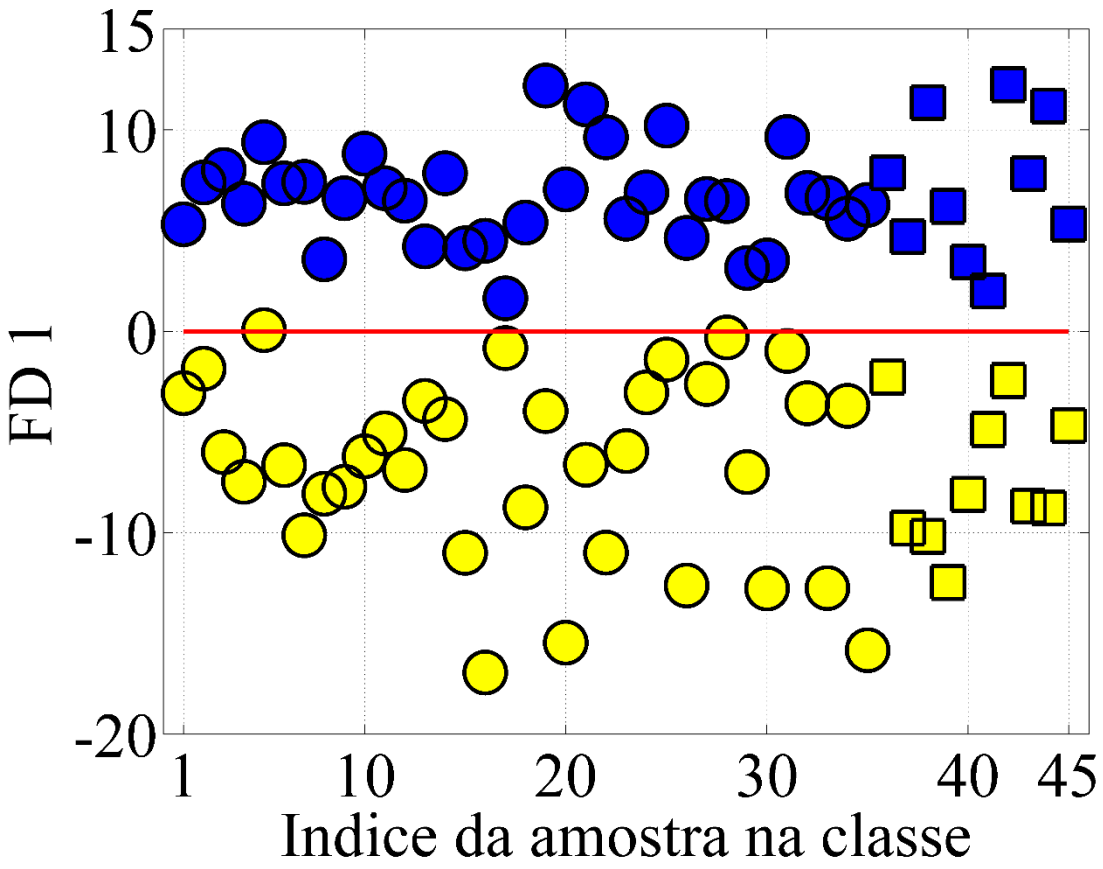

Figura 10 Gráficos de função discriminante de Fisher obtidos para PCA-GA-LDA, A linha vermelha indica o limite entre as classes, enquanto círculos e quadrados representam as amostras nos conjuntos de treinamento e teste, respectivamente, sendo azul para a classe 1 e amarelo para a classe 2.

Em relação ao modelo que utilizou o STEPWISE como técnica de seleção de componentes PCA-SW-LDA (**Figura 11**), este selecionou três PCs (PC3, PC6 e PC17). Apesar de ainda incluir duas componentes que não possuem informação discriminante, a técnica ainda apresentou um bom resultado, classificando corretamente todas as amostras nos conjuntos de treinamento e teste. Comparando os modelos baseados em seleção de variáveis por SW e GA, é possível observar que PC3 e PC6 foram incluídas nos dois modelos. Além disso, a melhoria do SW em relação ao GA deve-se à inclusão de um número reduzido de PCs não discriminativas, sendo menos prejudicada pela influência dessas componentes.

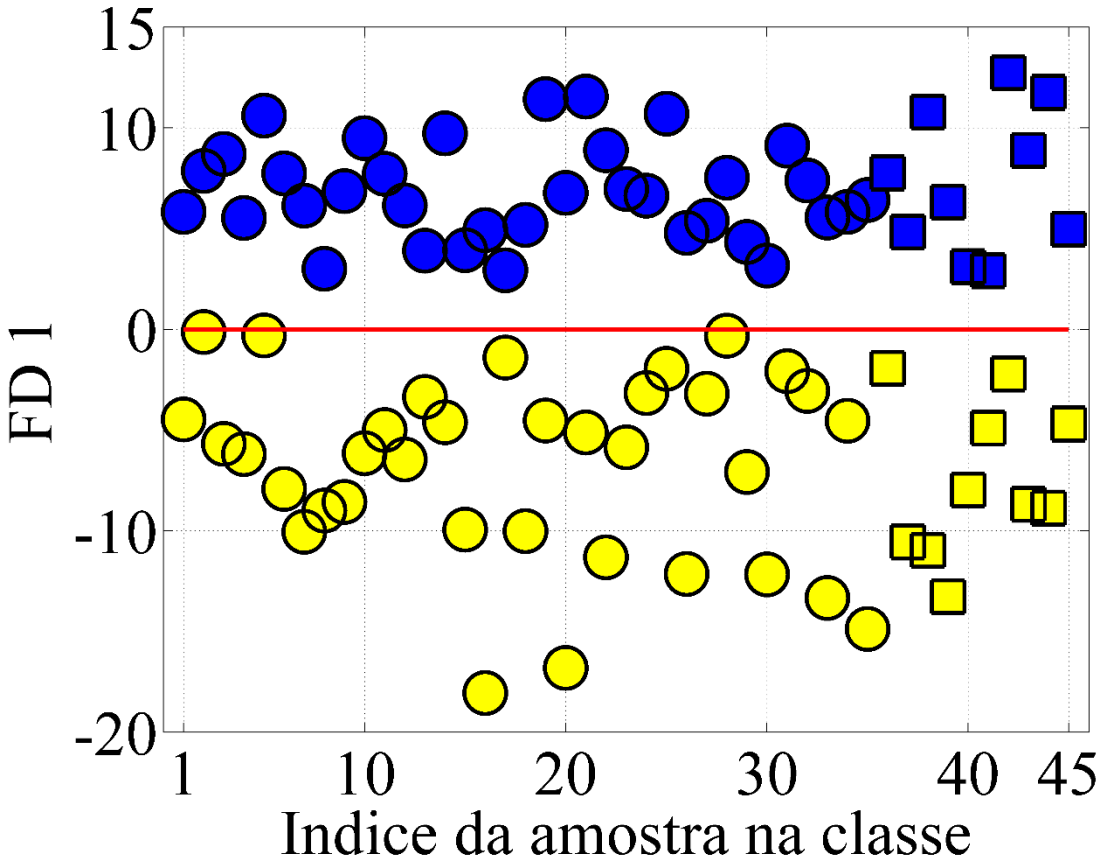

Figura 11 Gráficos de função discriminante de Fisher obtidos para PCA-SW-LDA, A linha vermelha indica o limite entre as classes, enquanto círculos e quadrados representam as amostras nos conjuntos de treinamento e teste, respectivamente, sendo azul para a classe 1 e amarelo para a classe 2.

Para o algoritmo PCA-DISP-LDA (**Figura 12**), como esperado, apenas a PC6 foi incluída no modelo, atingindo um TCC de 100% nos conjuntos de treinamento e teste. Apesar de apresentar o mesmo desempenho numérico do modelo PCA-SW-LDA, o PCA-DISP-LDA apresentou vantagem computacional e consequentemente sua parcimônia e interpretabilidade devem ser destacadas. Além disso, como o SW incluiu mais PCs que o necessário no modelo LDA, a classificação de amostras desconhecidas com matrizes mais complexas pode ser um fator limitante que leva a aumentar o número de erros de classificação em futuras aplicações analíticas.

Com base nos resultados apresentados aqui, é possível afirmar que o método proposto soluciona o problema de classificação de forma otimizada, construindo um modelo exatamente com as componentes necessárias e com o mínimo esforço computacional, tanto em termos do

modelo final gerado, quanto na identificação da quantidade e de quais componentes eram necessárias para a composição do modelo final de classificação.

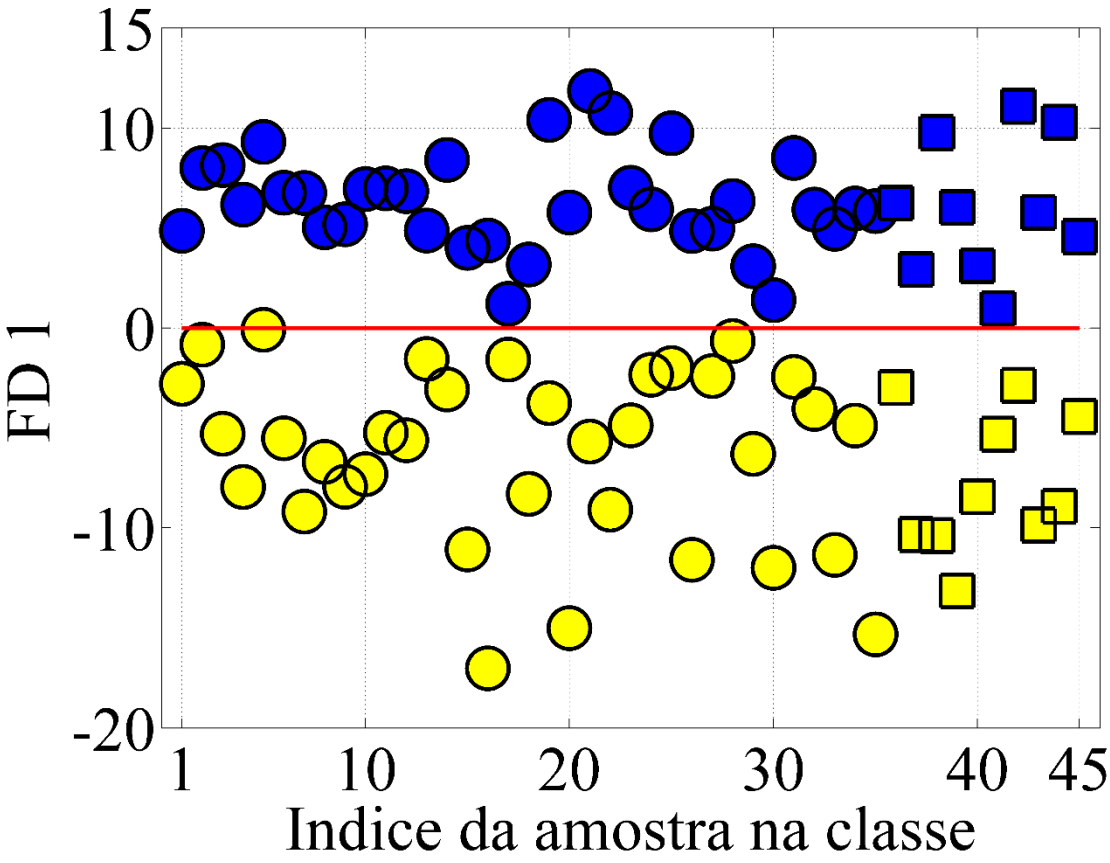

Figura 12 Gráficos de função discriminante de Fisher obtidos para PCA-DISP-LDA usando os dados simulados. A linha vermelha indica o limite entre as classes, enquanto círculos e quadrados representam as amostras nos conjuntos de treinamento e teste, respectivamente, sendo azul para a classe 1 e amarelo para a classe 2.

Com base nestes resultados preliminares em um sistema controlado e simulado, e sabendo que o conjunto de dados em questão é uma representação teórica de um eventual problema químico, é necessário avaliar o desempenho do algoritmo proposto em problemas reais de classificação. Portanto, o PCA-DISP-LDA será aplicado para resolver dois problemas de classificação envolvendo estudos de caso para análise de alimentos e um estudo de caso relacionado a adulteração combustível a partir de óleo de soja comestível, conforme discutido a seguir.

5.2 CLASSIFICAÇÃO DE ÓLEOS DE SOJA EM RELAÇÃO À DATA DE VALIDADE POR ESPECTROSCOPIA NIR

Na **Figura 13** são apresentados os espectros NIR obtidos a partir de óleos de soja não expirados (em linhas lilás) e expirados (em linhas verdes). Além disso, o desvio padrão para cada variável espectral também é indicado (em barras cinza). Como pode ser visto, existe uma tendência visual de separação entre as duas classes estudadas na região de 1400 a 1600 nm, que está associada aos primeiros sobretons dos grupos funcionais RC2OR e RCOR diretamente envolvidos no processo de oxidação de óleos vegetais.

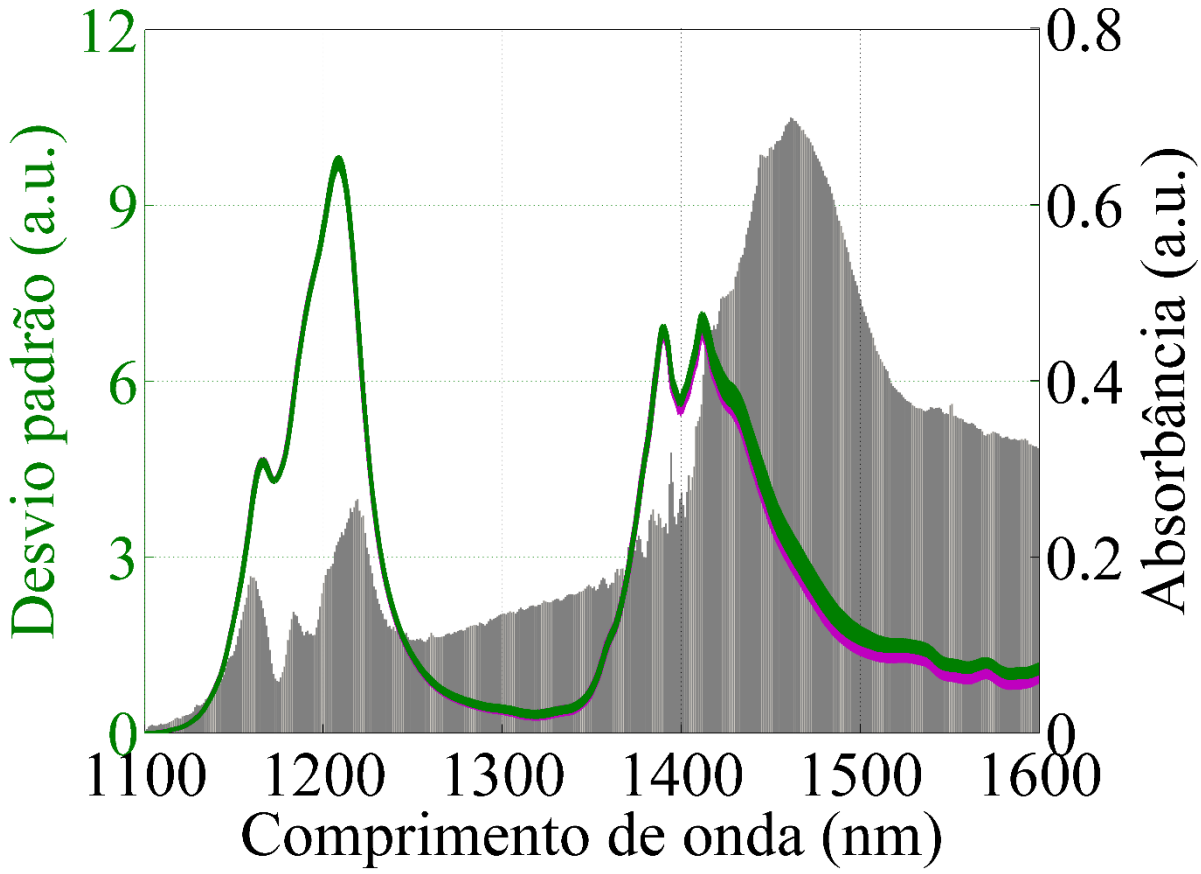

Figura 13 Espectros NIR obtidos na faixa de 1100 a 1600 nm a partir de óleos de não expirados (em lilás) e expirados (em verde). Os desvios padrões para cada variável são mostrados em barras cinza.

Após a PCA ser aplicada no espectro NIR, é possível observar nos gráficos apresentados na **Figura 14** que PC1 é responsável pela maior parte da variância explicada. Como consequência, PCA-LDA convencional tenderá a excluir PCs com menor variância explicada dos dados da construção do modelo. Além disso, os valores de discriminabilidade indicam que

PC1, PC9 e PC10 também podem conter informações necessárias para melhorar a classificação, em outras palavras, a parcimônia no PCA-LDA convencional, em termos de variáveis, prejudica a eficiência do modelo construído, exigindo uma etapa adicional para escolher os PCs mais relevantes.

Na **Tabela 2** são apresentados os resultados para a classificação dos óleos de soja em relação à data de vencimento por espectroscopia NIR usando PCA-LDA, PCA-GA-LDA, PCA-SW-LDA e PCA-DISP- LDA. Os scores e loadings de todas as componentes principais utilizadas nos modelos são mostrados no apêndice da seção 8.

Tabela 2. Resultados de classificação obtidos para PCA-LDA, PCA-GA-LDA, PCA- SW-LDA e PCA-DISP-LDA usando os espectros visíveis dos óleos de soja em relação à vida útil: não expirados e expirados.

Modelo	**PCs (Índice) incluídas no modelo LDA**	**TCC (%)**	
		Treinamento (erros)	**Teste (erros)**
PCA-LDA	2 (1,2)	91.43 (3)	86.67 (2)
PCA-GA-LDA	4 (1,6,9,10)	94.29 (2)	100 (0)
PCA-SW-LDA	4 (1,7,9,10)	94.29 (2)	100 (0)
PCA-DISP-LDA	2 (1,9)	97.14 (1)	100 (0)

Como esperado, o PCA-LDA convencional utilizou apenas PC1 e PC2, que representam 98% da variância explicada, atingindo taxas de classificação corretas de 91,43 e 86,67 para os conjuntos de treinamento e teste, respectivamente. As menores TCC para o PCA-LDA, podem ter ocorrido por PC9 não ter sido incluída ao modelo.

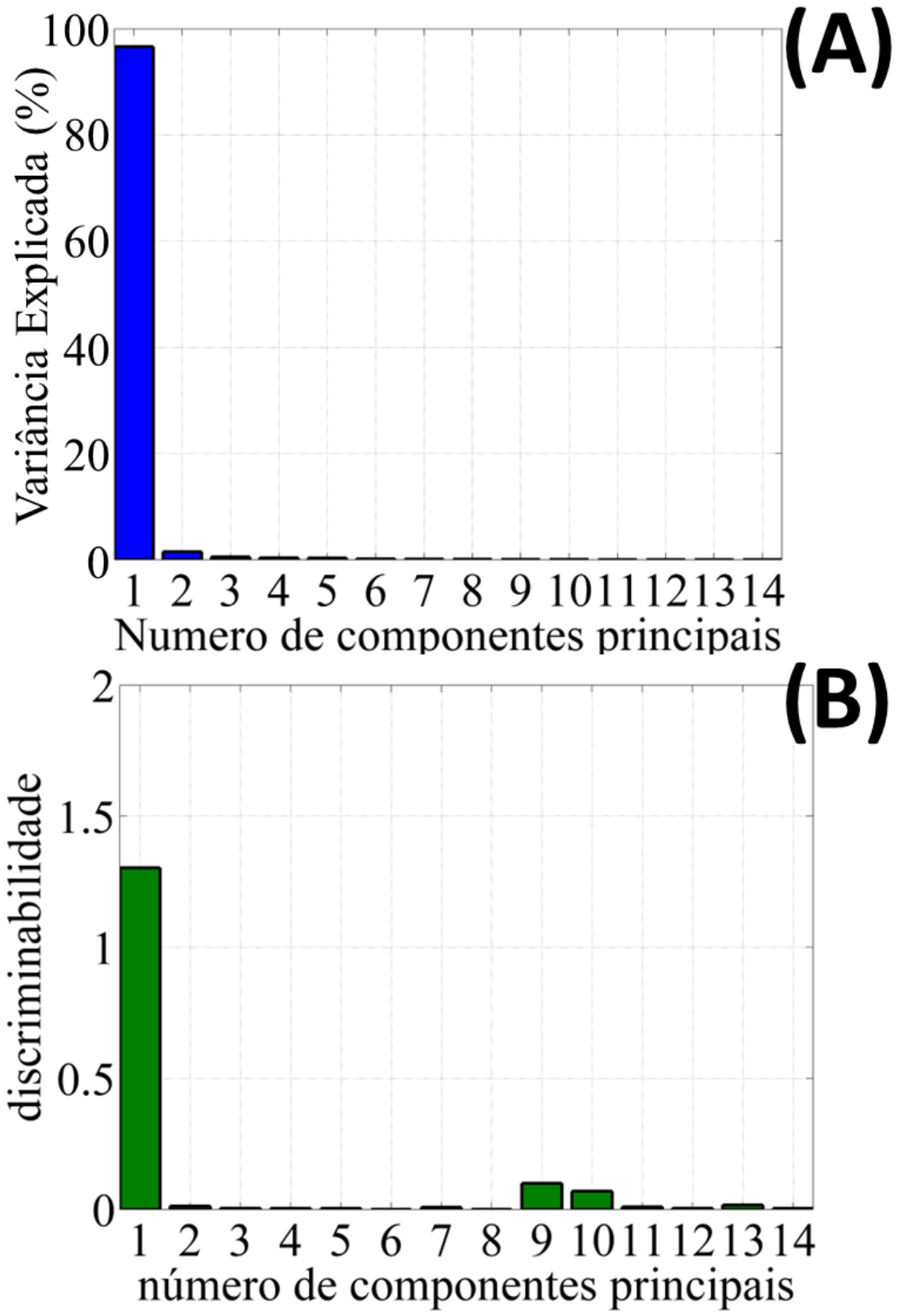

Figura 14 Valores de variância explicada versus discriminação de Fisher nas primeiras quatorze PCs.

Ainda de acordo com o que é apresentado na **Figura 14b** apesar da PC1 possuir alto valor de discriminabilidade a PC2 não contribui para a classificação das amostras, o que justifica o pior desempenho apresentado neste estudo de caso para o PCA-LDA. Além disso observou-se que a informação contida na componente principal 9, com baixíssima variância explicada, contribuiu na melhoria de classificação e generalização dos modelos, utilizaram essa PC.

Observa-se na **Figura 15** que três amostras foram classificadas incorretamente no conjunto de treinamento para ambas as classes. De igual forma para o conjunto de teste duas amostras, sendo uma de cada classe, são classificadas incorretamente, mostrando que o modelo pode não ser tão generalista quanto os demais modelos.

Esta informação corrobora com a premissa de que apenas a variância explicada não é suficiente para a determinação das componentes principais que são utilizadas na construção dos modelos PCA-LDA. É possível ainda verificar essa informação com base na distribuição das amostras em torno do limiar de classificação na **Figura 15**, no qual não há uma separação efetiva entre as duas classes, e há uma quantidade muito grande de amostras das duas classes próximas a linha que representa a fronteira traçada pelo modelo LDA, tanto para o conjunto de treinamento quanto para o conjunto de teste.

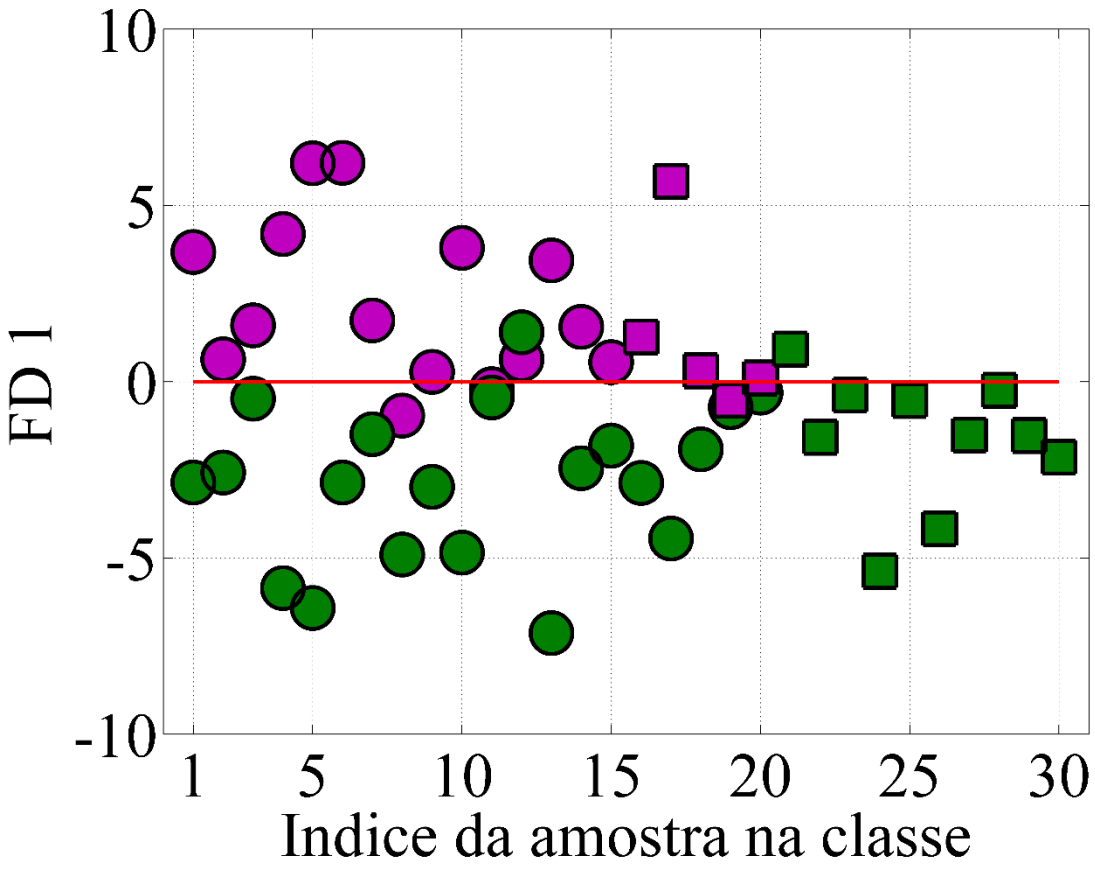

Figura 15 Gráfico da função discriminante de Fisher obtidos para PCA-LDA, A linha vermelha indica o limite entre as classes, não expirados (em lilás) e expirados (em verde), enquanto círculos e quadrados representam as amostras nos conjuntos de treinamento e teste, respectivamente.

Os modelos baseados em GA (**Figura 16**) e SW (**Figura 17**) selecionaram quatro componentes, sendo as três PCs com os maiores valores de discriminabilidade de Fisher (PC1, PC9 e PC10), além da inclusão da PC6 para GA e PC7 para SW. Nesse caso, ambos os modelos alcançaram as mesmas taxas de classificação corretas para os conjuntos de treinamento (94,29%) e teste (100%), com duas classificações incorretas (uma de cada classe) no conjunto de treinamento, demonstrando que PC6 e PC7 influenciaram negativamente para o desempenho dos modelos, quando comparados com a classificação do modelo PCA-DISP-LDA.

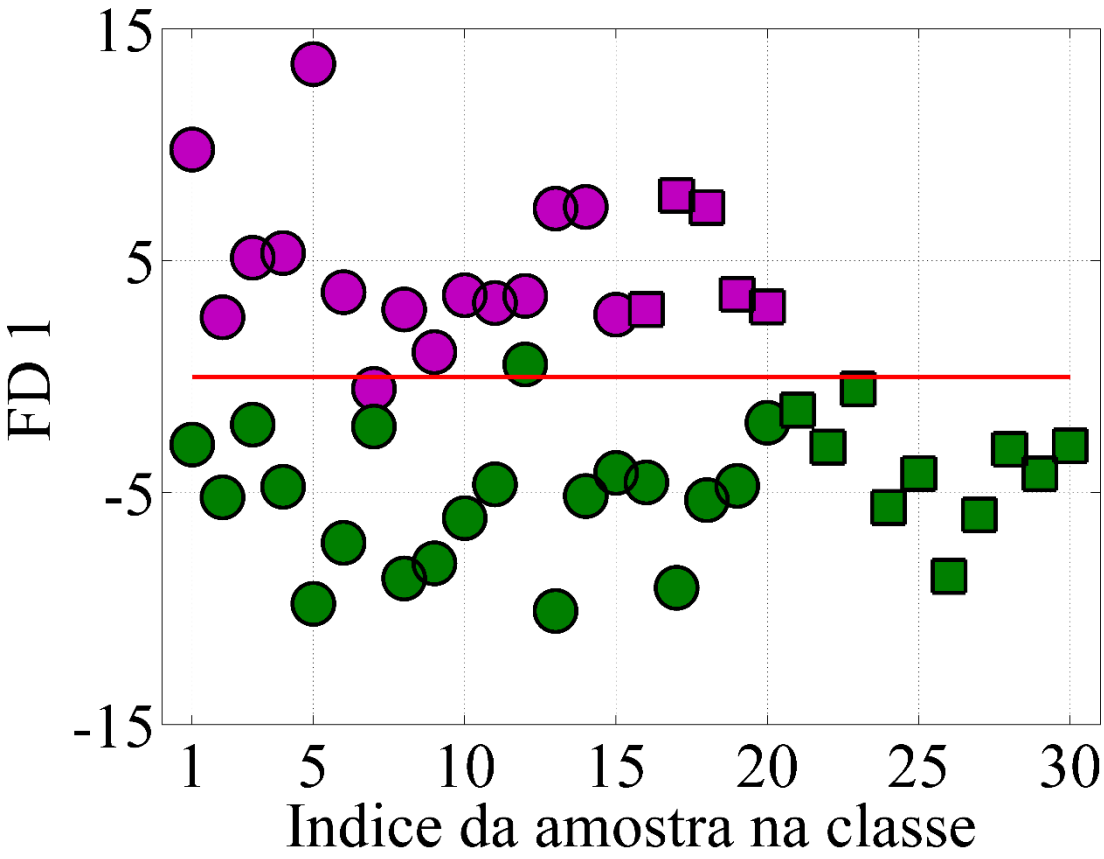

Figura 16 Gráfico da função discriminante de Fisher obtidos para PCA-GA- LDA, A linha vermelha indica o limite entre as classes, não expirados (em lilás) e expirados (em verde), enquanto círculos e quadrados representam as amostras nos conjuntos de treinamento e teste, respectivamente.

Apesar de ter melhores resultados do que os PCA-LDA convencionais, os modelos baseados em GA e SW exigem cálculos mais complexos para a escolha das PCs, além de um analista hábil para otimizar e ajustar alguns parâmetros de seleção de variáveis, como o número da população e as gerações para GA e o limiar de correlação para SW.

É importante ainda que os modelos GA e SW utilizaram a PC10. Embora esta componente principal tenha o terceiro maior valor de discriminabilidade, ela não foi suficiente para a melhorar a classificação. Em testes realizados, a inclusão da PC10 ao modelo DISP também não apresentou alteração no desempenho de classificação, e com base na comparação do modelo DISP com os modelos GA e SW, a influência negativa das PCs 6 e 7 superam a influência positiva da PC10, vide os resultados de classificação.

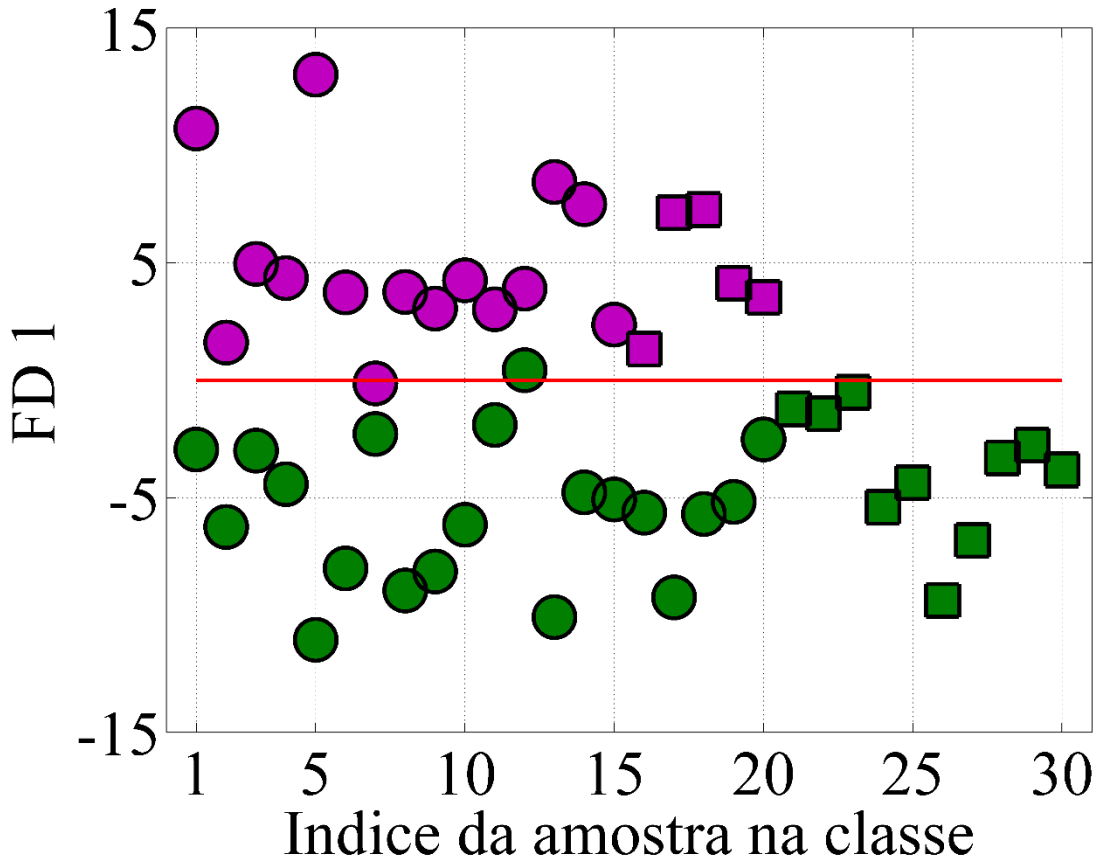

Figura 17 Gráfico da função discriminante de Fisher obtidos para PCA-SW-LDA, A linha vermelha indica o limite entre as classes, não expirados (em lilás) e expirados (em verde), enquanto círculos e quadrados representam as amostras nos conjuntos de treinamento e teste, respectivamente.

O modelo de classificação PCA-DISP-LDA (**Figura 18**) foi o mais parcimonioso neste estudo de caso em termos da quantidade de PCs utilizadas (PC1 e PC9). Este comportamento pode ser explicado pela etapa de reordenação das PCs em termos de seus valores de discriminabilidade, que fornece ao modelo LDA as componentes principais mais discriminantes.

Em contraponto ao que ocorreu no modelo PCA-LDA, as PCs 2 a 8 não precisariam ser incluídas no modelo para que a PC9 compusesse o modelo de classificação, otimizando o modelo construído e simplificando a escolha das componentes principais.

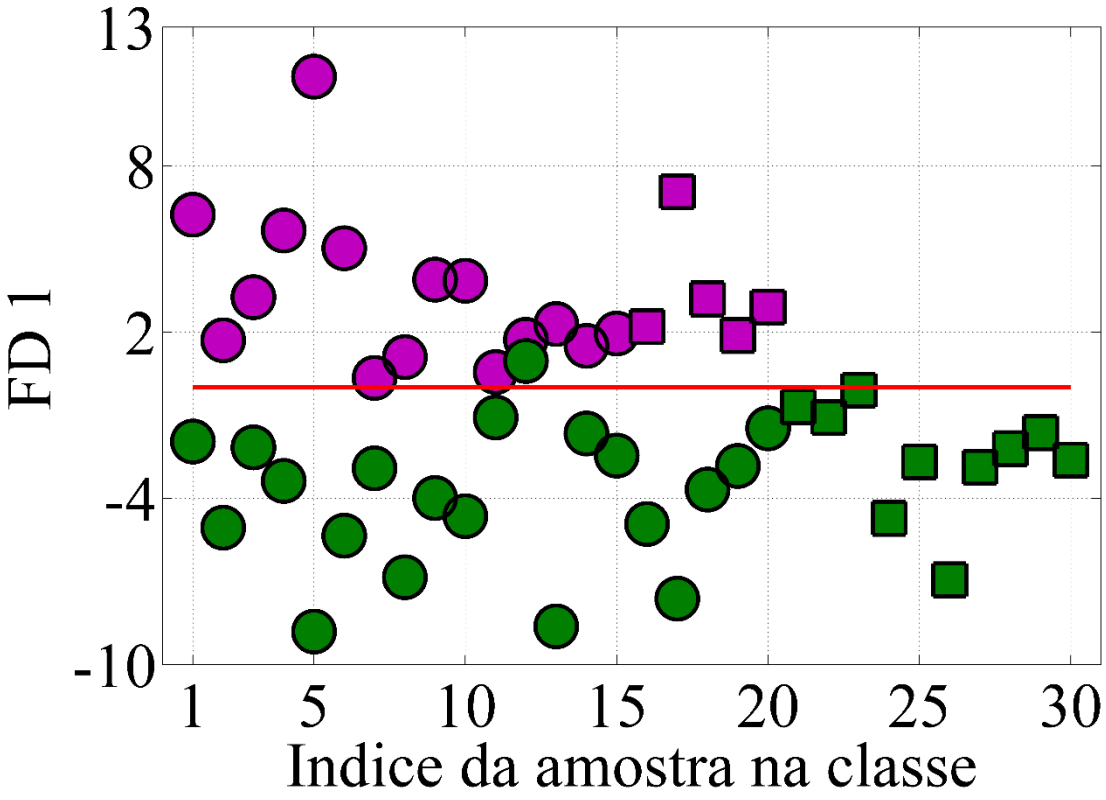

Figura 18 Gráfico da função discriminante de Fisher obtidos para PCA-DISP-LDA, A linha vermelha indica o limite entre as classes, não expirados (em lilás) e expirados (em verde), enquanto círculos e quadrados representam as amostras nos conjuntos de treinamento e teste, respectivamente.

Na **Figura 18** é apresentado o gráfico da função discriminante de Fisher para o modelo PCA-DISP-LDA. Apenas uma amostra foi classificada incorretamente no conjunto de treinamento, proporcionando taxas de classificação corretas de 97,14 e 100% para os conjuntos de treinamento e teste, respectivamente.

É possível verificar que poucas amostras estão dispostas no limite da linha que estabelece o limiar entre as classes, diferente do que ocorre no modelo PCA-LDA.

5.3 IDENTIFICAÇÃO DA ADULTERAÇÃO DE MISTURAS DE BIODIESEL/DIESEL COM ÓLEO DE SOJA PORESPECTROSCOPIA VIS-NIR

Os espectros Vis-NIR das amostras de mistura de biodiesel / diesel adulterado e não adulterado são ilustrados na **Figura 19**.

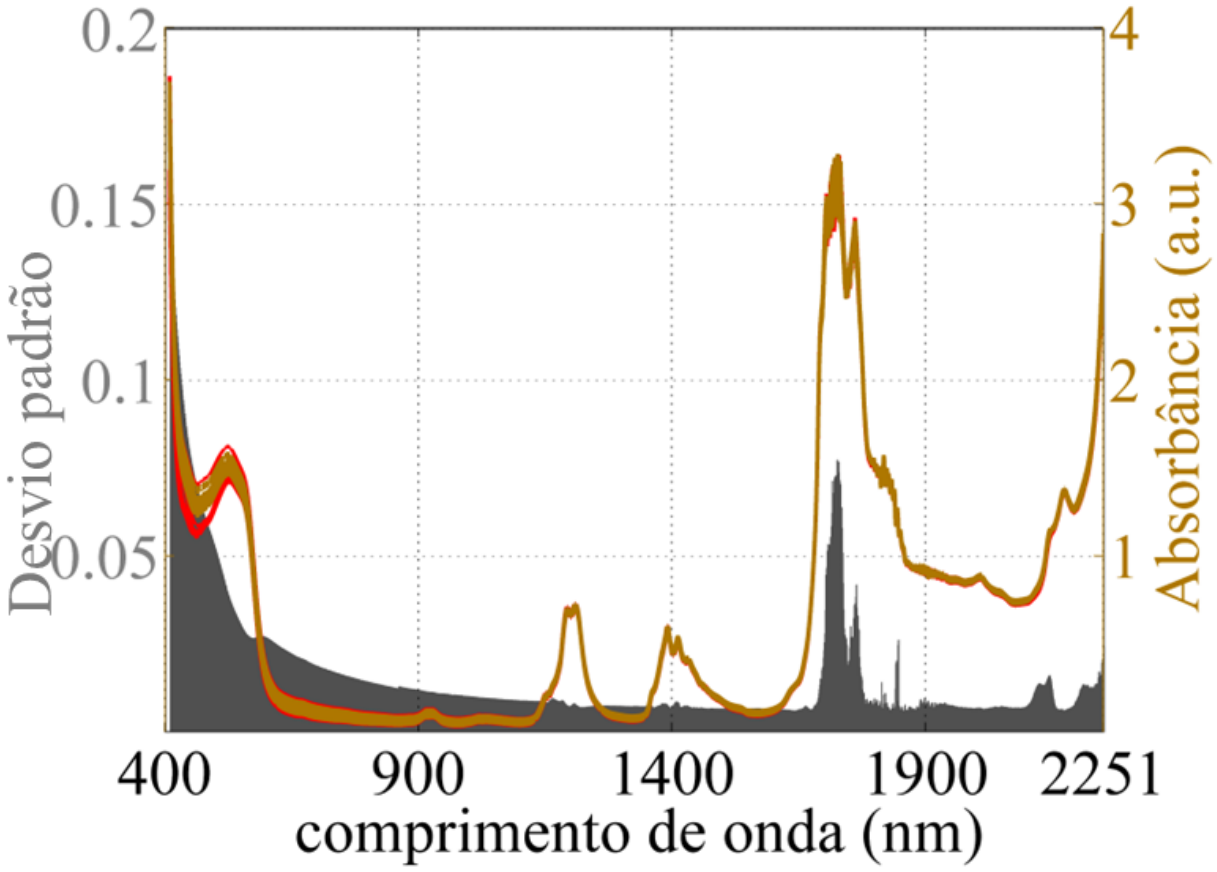

Figura 19 Espectros Vis-NIR obtidos na faixa de 410 a 2251 nm a partir de misturas de biodiesel/diesel B5: não adulteradas (em vermelho) e adulteradas com óleo de soja na faixa de 0,5 a 2,5% (v / v) (em marrom). Os desvios padrões para cada variável são mostrados em barras cinza.

Na **Figura 20b** é demonstrado que PC1, PC11, PC17, PC6, PC10 e PC13 têm os mais altos valores de discriminabilidade em ordem decrescente, o que indica que esses componentes podem ser os responsáveis por melhorias nos modelos de classificação, embora a mesma inferência não seja possível ao verificar-se as duas primeiras PCs (**Figura 20a**), que são responsáveis pela maior parte da variância explicada, conforme sugerido pelo modelo PCA-LDA.

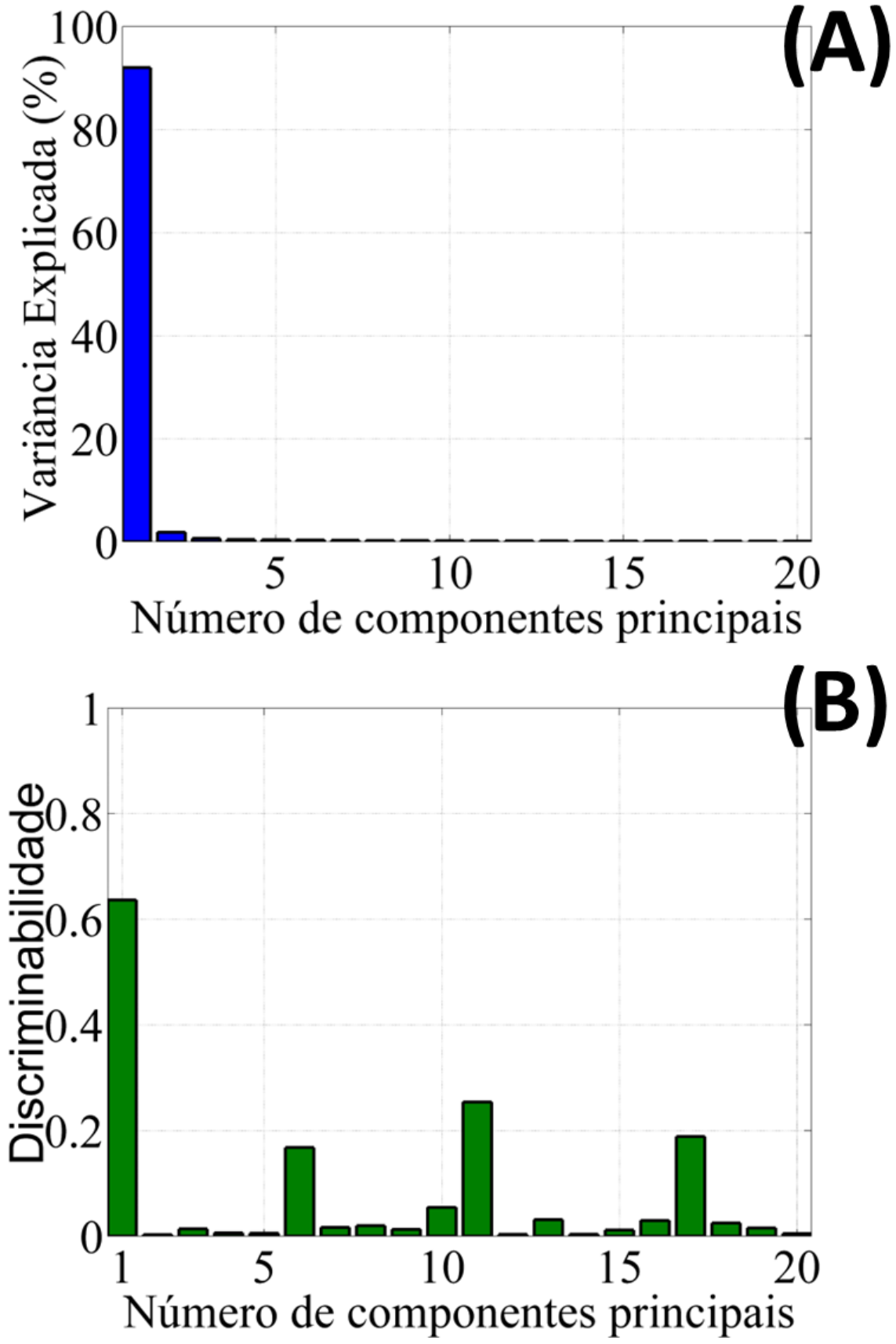

Figura 20 (a) Valores de variância explicada e (b) discriminabilidade de Fisher para as primeiras vinte PCs.

Os resultados da classificação obtidos por PCA-LDA, PCA-GA-LDA, PCA-SW- LDA e PCA-DISP-LDA são apresentados na **Tabela 3** e também podem ser visualizados nos gráficos das funções discriminantes de Fisher nas **Figuras 22** a **24**. Os scores e loadings de todas as componentes principais utilizadas nos modelos são mostrados no apêndice da seção 8 ao final do documento.

Tabela 3. Resultados da classificação obtidos para PCA-LDA, PCA-GA-LDA, PCA- SW-LDA e PCA-DISP-LDA usando os espectros Vis-NIR das misturas de biodiesel / diesel B5: não adulterados e adulterados com óleo de soja na faixa de 0,5 a 2,5% (v / v).

Modelo	**PCs (Índice) incluídas no modelo LDA**	**TCC (%)**	
		Treinamento (erros)	**Teste (erros)**
PCA-LDA	7 (1,2,3,4,5,6,7)	91.94 (5)	92.86 (2)
PCA-GA-LDA	7 (1,6,7,10,11,13,17)	100 (0)	100 (0)
PCA-SW-LDA	4 (1,10,11,13)	91.94 (5)	100 (0)
PCA-DISP-LDA	6 (1,6,10,11,13,17)	100 (0)	100 (0)

O PCA-LDA convencional obteve 91,94 e 92,86% de taxa de classificação correta para os conjuntos de treinamento e teste, respectivamente. Neste modelo as sete primeiras PCs, que incluem pelo menos as PC1 e PC6, com altos valores de discriminabilidade foram usados.
Para este modelo, apenas amostras não adulteradas foram classificadas incorretamente como amostras adulteradas, sendo cinco amostras no conjunto de treinamento e duas no conjunto de teste. Todas as amostras adulteradas foram classificadas corretamente em suas respectivas classes, como mostra a **Figura 21**.

Vale destacar aqui, que é ideal, que amostras adulteradas sejam classificadas corretamente, evitando que futuramente, fraudes não fossem identificadas por erros de classificação.

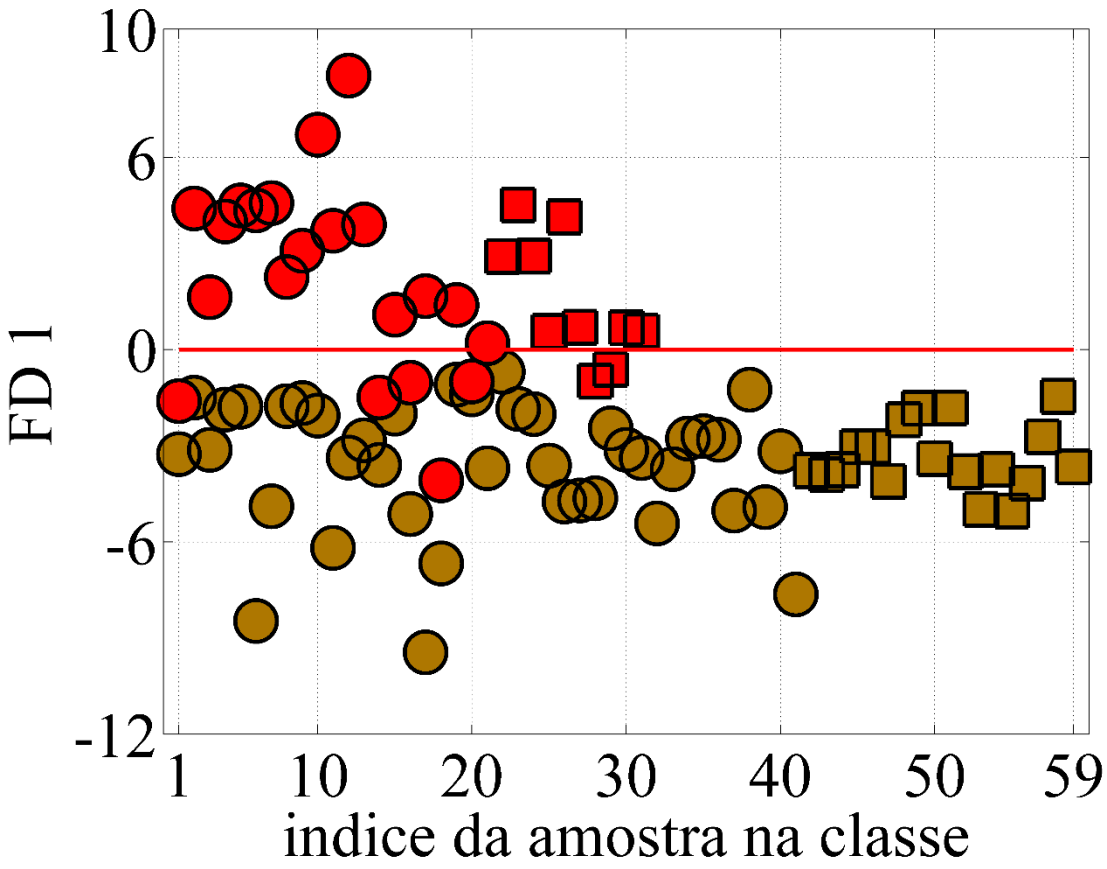

Figura 21 Os gráficos de funções discriminantes de Fisher obtidos para PCA-LDA usando os espectros Vis-NIR, das amostras não adulteradas (em vermelho) e adulteradas com óleo de soja na faixa de 0,5 a 2,5% (v / v) (em marrom). A linha vermelha indica o limite entre as classes, enquanto círculos e quadrados representam as amostras nos conjuntos de treinamento e teste, respectivamente.

Quando o SW é aplicado, o desempenho do modelo foi aprimorado, especialmente no conjunto de teste em que todas as amostras foram classificadas corretamente. Novamente, cinco amostras não adulteradas foram classificadas de forma incorreta como amostras adulteradas no conjunto de treinamento, atingindo uma taxa de classificação correta de 91,94% (**Figura 22**). Essa diminuição na taxa de erro no conjunto de testes deve-se à seleção de PCs com maiores valores de discriminabilidade, como PC1, PC10, PC11 e PC13. Porém, ainda falha ao não adicionar as componentes 6 e 17, que tem relevante poder discriminante.

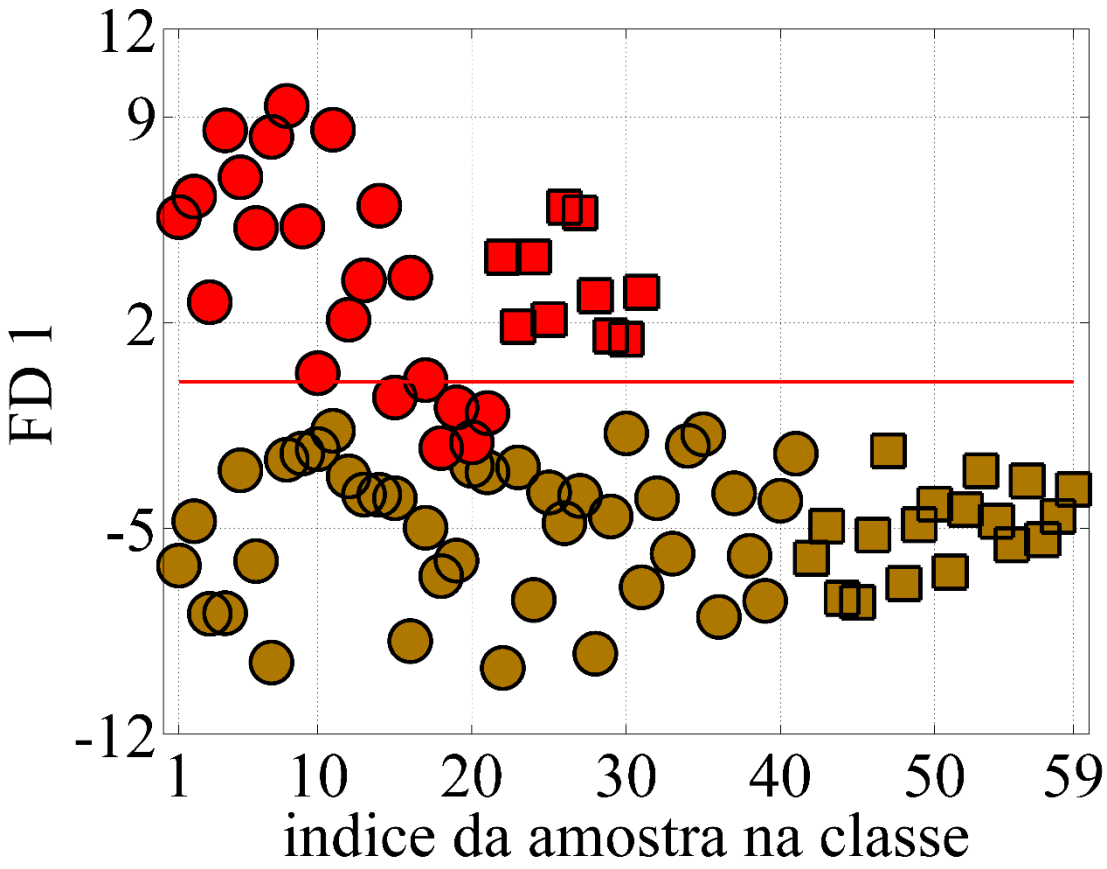

Figura 22 Os gráficos de funções discriminantes de Fisher obtidos para PCA-SW-LDA usando os espectros Vis-NIR, das amostras não adulteradas (em vermelho) e adulteradas com óleo de soja na faixa de 0,5 a 2,5% (v / v) (em marrom). A linha vermelha indica o limite entre as classes, enquanto círculos e quadrados representam as amostras nos conjuntos de treinamento e teste, respectivamente.

Por outro lado, PCA-GA-LDA e PCA-DISP-LDA classificaram corretamente todas as amostras nos conjuntos de treinamento e teste (**Figura 23 e 24**), selecionando os mesmos componentes do PCA-SW-LDA (PC1, PC10, PC11 e PC13), além de PC6, PC7 e PC17 no caso do GA e PC6 e PC17 no caso do DISP. Assim, podemos atribuir facilmente a melhoria da TCC no conjunto de treinamento a PC6 e PC17, demonstrando que o DISP foi novamente mais parcimonioso, uma vez que não adicionou a PC7 que apresenta baixo valor de discriminabilidade (**Figura 20**).

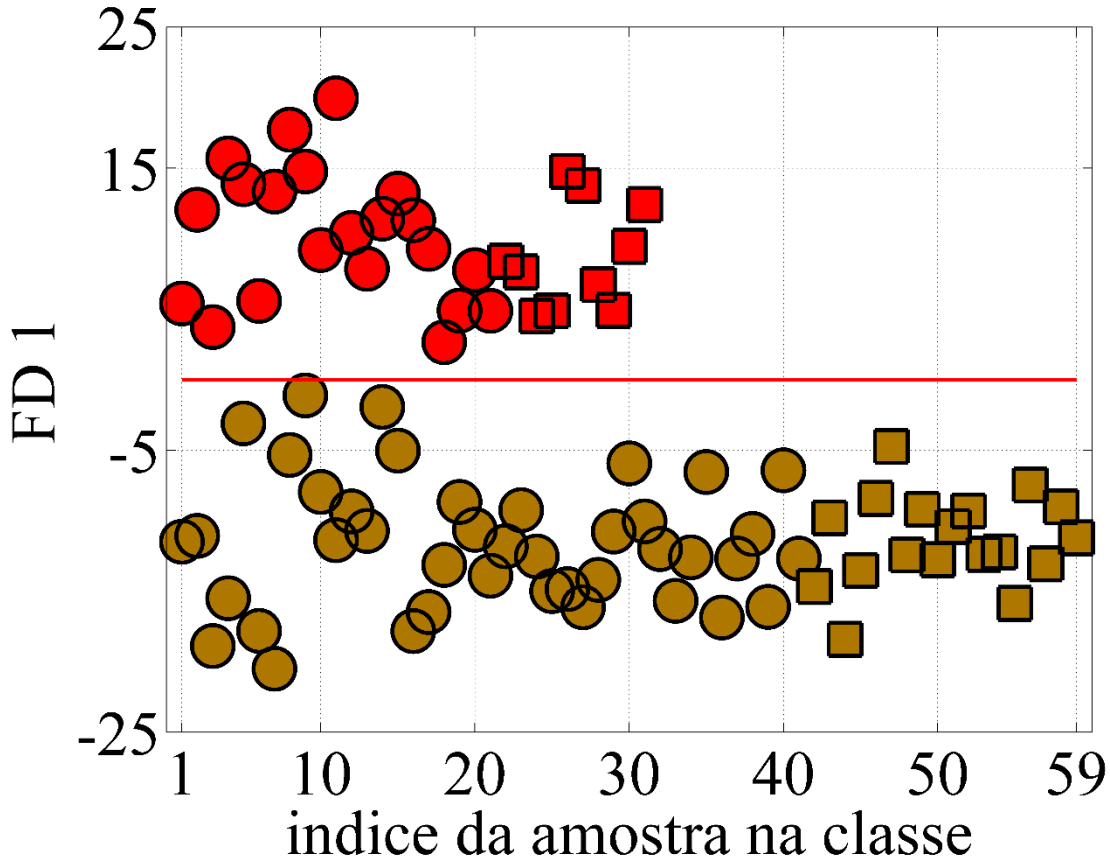

Figura 23 Os gráficos de funções discriminantes de Fisher obtidos para PCA-GA-LDA usando os espectros Vis-NIR, das amostras não adulteradas (em vermelho) e adulteradas com óleo de soja na faixa de 0,5 a 2,5% (v / v) (em marrom). A linha vermelha indica o limite entre as classes, enquanto círculos e quadrados representam as amostras nos conjuntos de treinamento e teste, respectivamente.

Embora numericamente os desempenhos de classificação do modelo PCA-DISP-LDA e do modelo PCA-GA-LDA sejam equivalentes, já que obtiveram as mesmas taxas de classificação correta, o modelo DISP foi mais parcimonioso tanto em termos de número de componentes escolhidos quanto no esforço computacional que é requerido para a escolha das PCs pelo algoritmo genético, além disso o PCA-GA-LDA requer uma série de ajuste nos parâmetros de entrada, diferentemente do que ocorre no DISP, que não precisa de nenhum tipo de ajuste de parâmetros de entrada e a matemática envolvida é bastante simples, o que torna sua utilização rápida e fácil de ser executada, permitindo inclusive que seja implementado diretamente no algoritmo de geração das PCs sem nenhuma manipulação por parte do analista, o que não é possível no caso do PCA-GA-LDA.

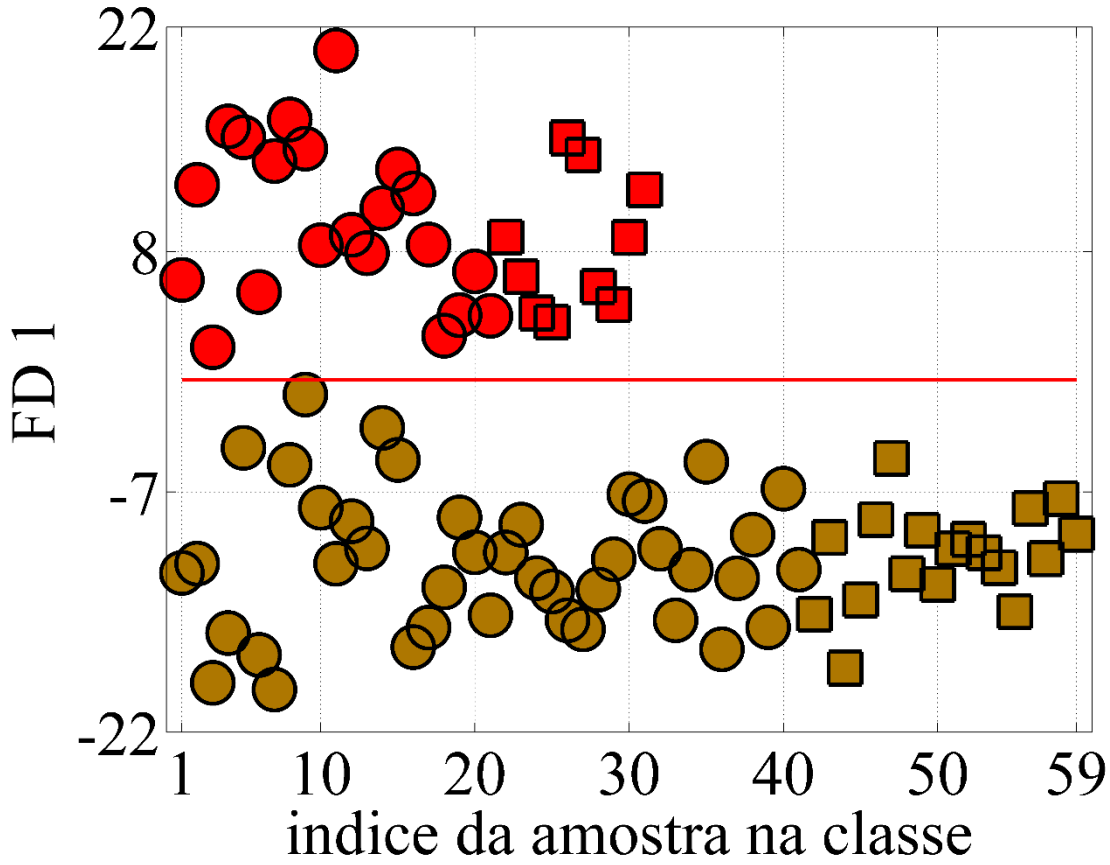

Figura 24 Os gráficos de funções discriminantes de Fisher obtidos para PCA-DISP-LDA usando os espectros Vis-NIR, das amostras não adulteradas (em vermelho) e adulteradas com óleo de soja na faixa de 0,5 a 2,5% (v / v) (em marrom). A linha vermelha indica o limite entre as classes, enquanto círculos e quadrados representam as amostras nos conjuntos de treinamento e teste, respectivamente.

Neste estudo de caso mais uma vez o método proposto demonstrou coerência nos resultados e possibilitou uma melhora de 7,8% no número total de classificações corretas quando comparadas à abordagem convencional, além de apresentar desempenho superior ou equivalente a estratégias já bem estabelecidas na literatura como PCA-GA-LDA e PCA-SW-LDA.

5.4 IDENTIFICAÇÃO DA ADULTERAÇÃO DE MISTURAS DE BIODIESEL/DIESEL COM ÓLEO DE SOJA PORESPECTROSCOPIA VIS-NIR

Os espectros NIR das amostras de cachaça adulteradas (em cinza) e não adulteradas (em azul) bem como o desvio padrão de cada variável em barra cinza estão ilustradas na **Figura 25**.

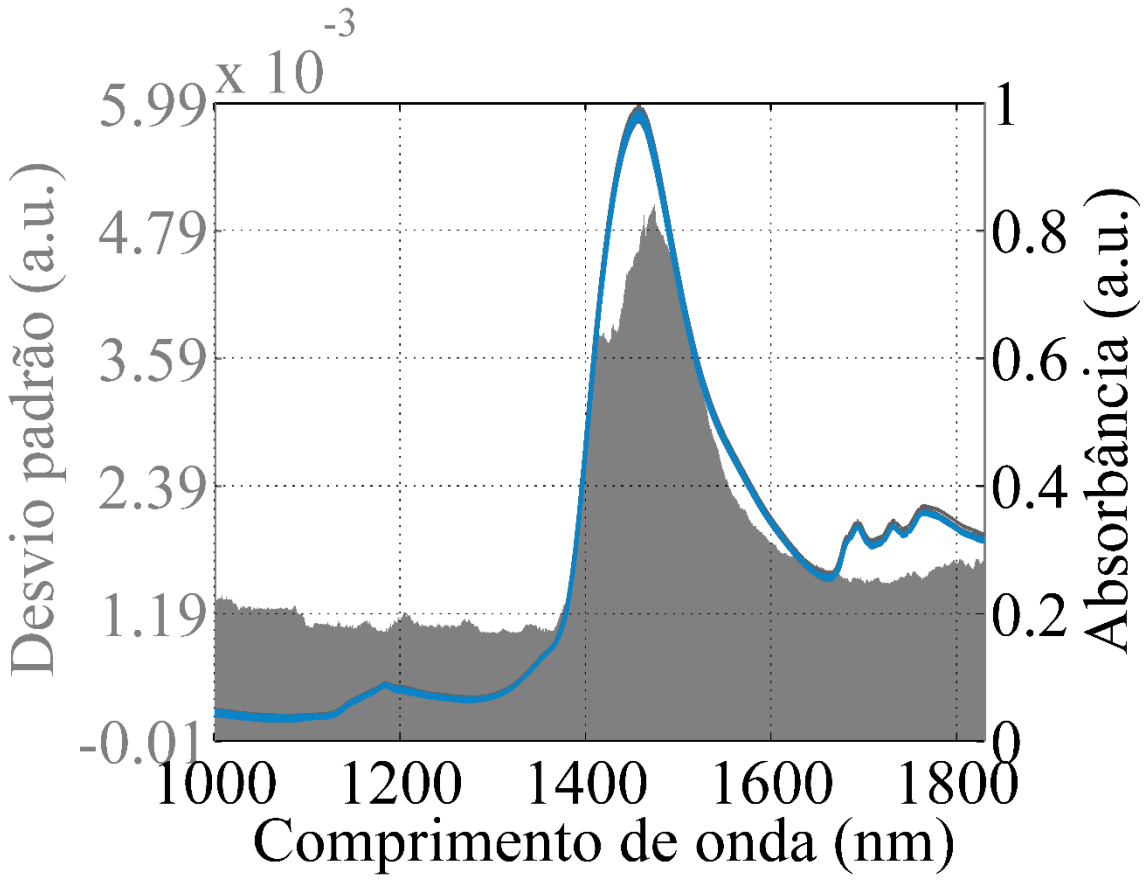

Figura 25 Espectros NIR obtidos na faixa de 1000 a 1600 nm a partir de amostras de cachaça envelhecidas (em cinza) e adulteradas com extrato de madeira (em azul). Os desvios padrões para cada variável são mostrados em barras cinza.

É possível verificar que os maiores valores de desvio padrão estão entre a região de 1400 e 1600nm. De acordo com a figura, ainda se nota que não há regiões de clara separação entre as classes, sendo praticamente todo o espectro sobreposto. Na **Figura 26** são mostrados os valores de variância explicada e discriminabilidade de Fisher para as PCs calculadas para este estudo de caso.

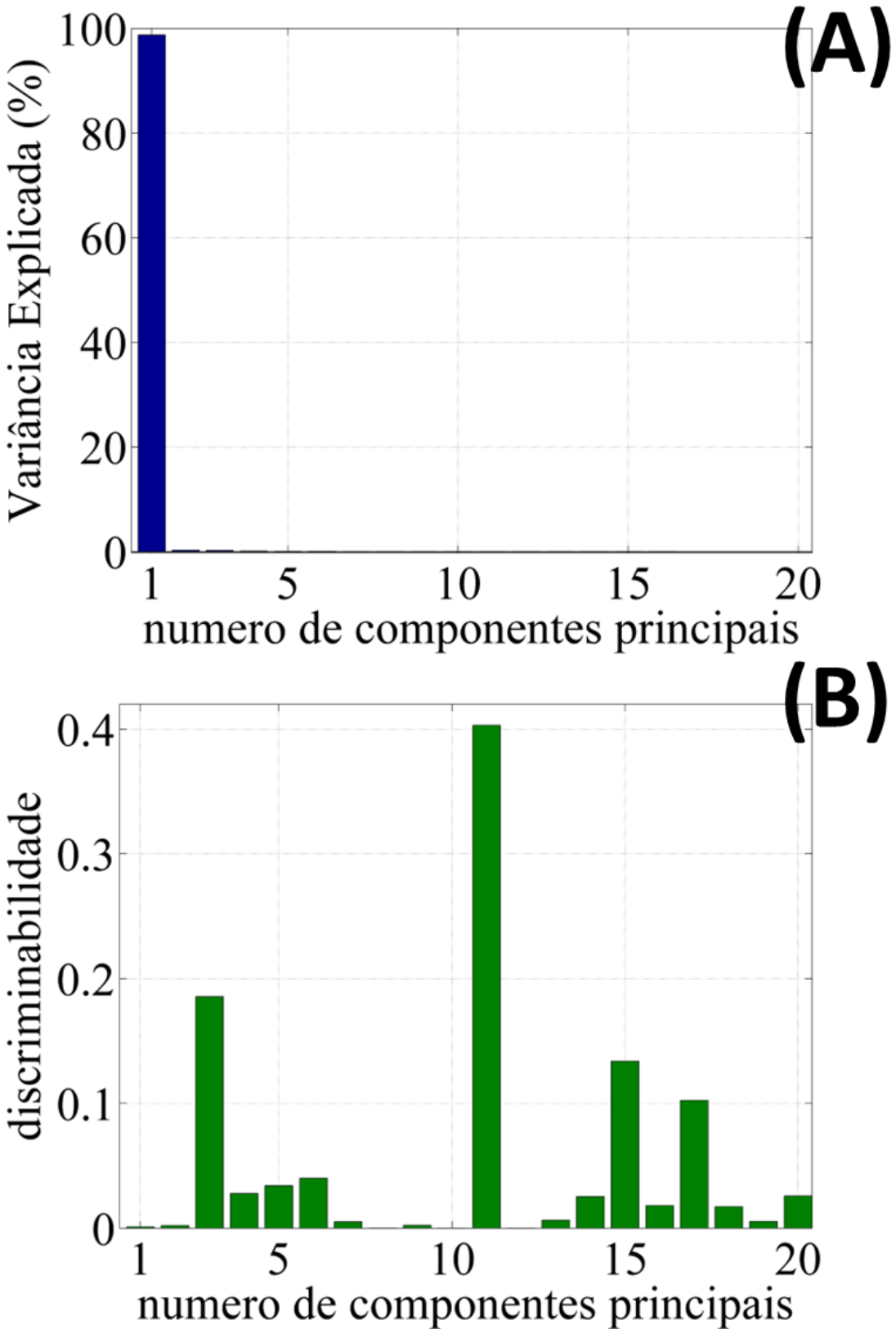

Figura 26 Valores de variância explicada versus discriminação de Fisher para as primeiras vinte PCs.

Apesar de apresentar 99% da variância explicada total dos dados como demonstrado na **Figura 26a** as componentes 1 e 2 possuem poder discriminante muito menores que os observados para a PC 11, como verificado na **Figura 26b**. Em contrapartida a PC3 que tem apenas 0,2% de variância explicada possui o segundo maior poder discriminante, assim como a PC11 que tem 0,02% de variância explicada, mas tem o maior poder discriminante deste estudo de caso. Os scores e loadings de todas as componentes principais utilizadas nos modelos são mostrados no apêndice da seção 8 ao final do documento.

Juntas, as 8 PCs (3,4,5,6,11,15,17,20) que apresentam os maiores valores de discriminabilidade de Fisher, possuem apenas 0.5% de variância explicada. É previsível que o PCA-LDA tenha dificuldade em obter bons resultados neste cenário, que diferente dos bancos de dados experimentais anteriores, não possui poder discriminante relevante nas primeiras componentes principais. Em adição, PCs ranqueadas em alta posição em termos de variância explicadas possuem os maiores poderes discriminantes, desta forma, o modelo PCA-LDA necessitaria adicionar uma quantidade muito grande de variáveis para adicionar a discriminabilidade presentes nesta PCs.

Na **Tabela 4** é possível verificar o desempenho dos modelos de classificação. Como esperado, o modelo PCA-LDA possui um desempenho baixo em termos de taxa de classificação correta. Um outro ponto de mínimo é verificado quando o modelo é construído com 16 PCs, porém, claramente trata-se de um modelo sobre ajustado, levando em conta a relação entre a quantidade de componentes e de amostras no conjunto de treinamento e teste para este estudo.

Na **Figura 27** é apresentado o gráfico da função discriminante para o modelo PCA-LDA convencional. É possível identificar uma grande quantidade de erros de classificação tanto no conjunto de treinamento quanto no conjunto de teste, e além disso, de modo geral, as amostras estão distribuídas bem próximas ao limiar de classe, demonstrando que o PCA-LDA não consegui obter bons resultados neste cenário, apresentando 14 erros no conjunto de treinamento e 4 no conjunto de teste.

Tabela 4. Resultados da classificação obtidos para PCA-LDA, PCA-SW-LDA, PCA-GA-LDA e PCA-DISP-LDA para amostras de cachaça envelhecidas e adulteradas com extrato de madeira.

Modelo	PCs (Índice) incluídas nomodelo LDA	TCC (%)	
		Treinamento (erros)	Teste (erros)
PCA-LDA	9 (1,2,3,4,5,6,7,8,9)	81,6 (14)	73,3 (4)
PCA-GA-LDA	7 (3,8,11,12,15,16,18)	97,4 (2)	93,3 (1)
PCA-SW-LDA	5 (3,6,9,11,17)	93,4 (5)	93,3 (1)
PCA-DISP-LDA	8 (3,4,5,6,11,15,17,20)	100 (0)	100 (0)

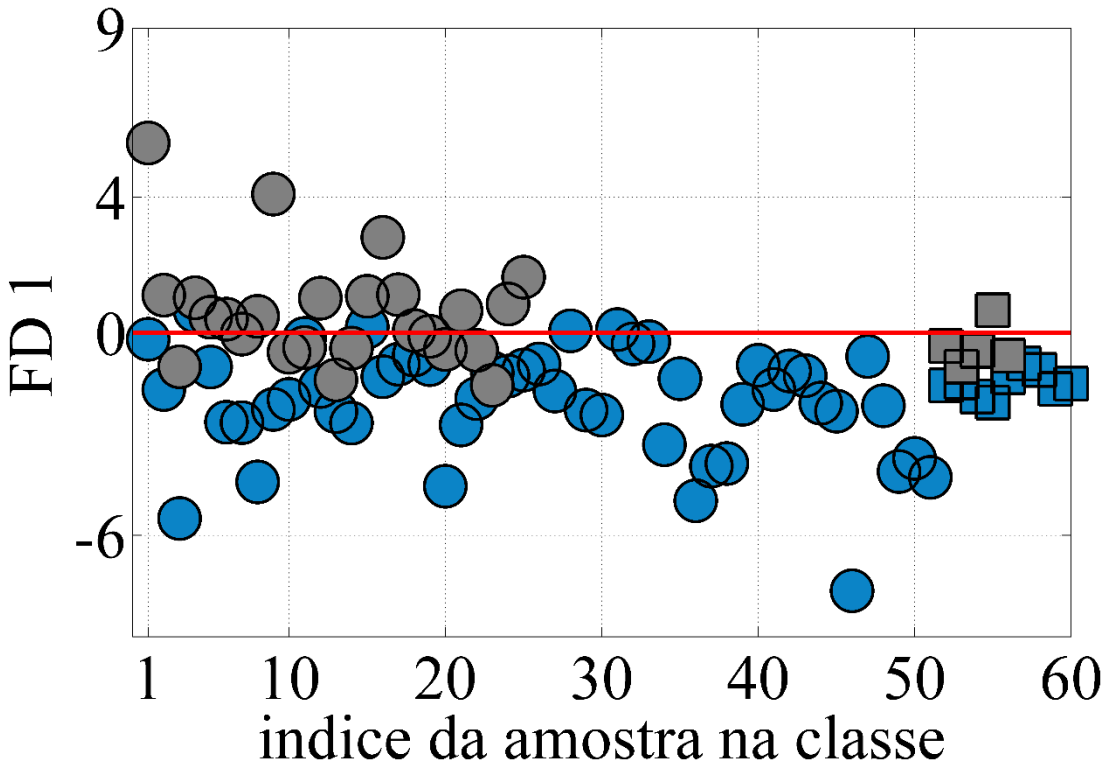

Figura 27 Os gráficos de funções discriminantes de Fisher obtidos para PCA-LDA usando os espectros NIR das amostras não adulteradas (em azul) e adulteradas com extratos de madeira (em cinza). A linha vermelha indica o limite entre as classes, enquanto círculos e quadrados representam as amostras nos conjuntos de treinamento e teste, respectivamente. amostras não adulteradas (em azul) e adulteradas com extratos de madeira (em cinza).

O modelo PCA-SW-LDA mostrou melhores resultados quando comparado ao modelo PCA-LDA, uma vez que não está limitado à utilização das componentes na ordem definida pela variância explicada. Isto permitiu que 5 PCs fossem selecionadas, dentre as quais se encontram as PCs 3, 6, 11, e 17, que estão entre os maiores poderes discriminantes, a PC 9 foi a única com baixo valor de discriminabilidade entre as 5.

Na **Figura 28** observa-se que amostras adulteradas e não adulteradas foram classificadas incorretamente no conjunto de treinamento, o que não é considerado um bom cenário, uma vez que amostras falsificadas poderiam ser vendidas de forma irregular e não seriam detectadas pelo modelo PCA-SW-LDA. Além disso uma amostra adulterada do conjunto de teste foi classificada como amostra não adulterada, assim como ocorreu no conjunto de treinamento. De modo geral o modelo PCA-SW-LDA apresentou 6 erros de classificação, sendo 5 no conjunto de treinamento e uma no conjunto de teste, perfazendo uma taxa de classificação correta de 93,4 e 93,3 para treinamento e teste respectivamente.

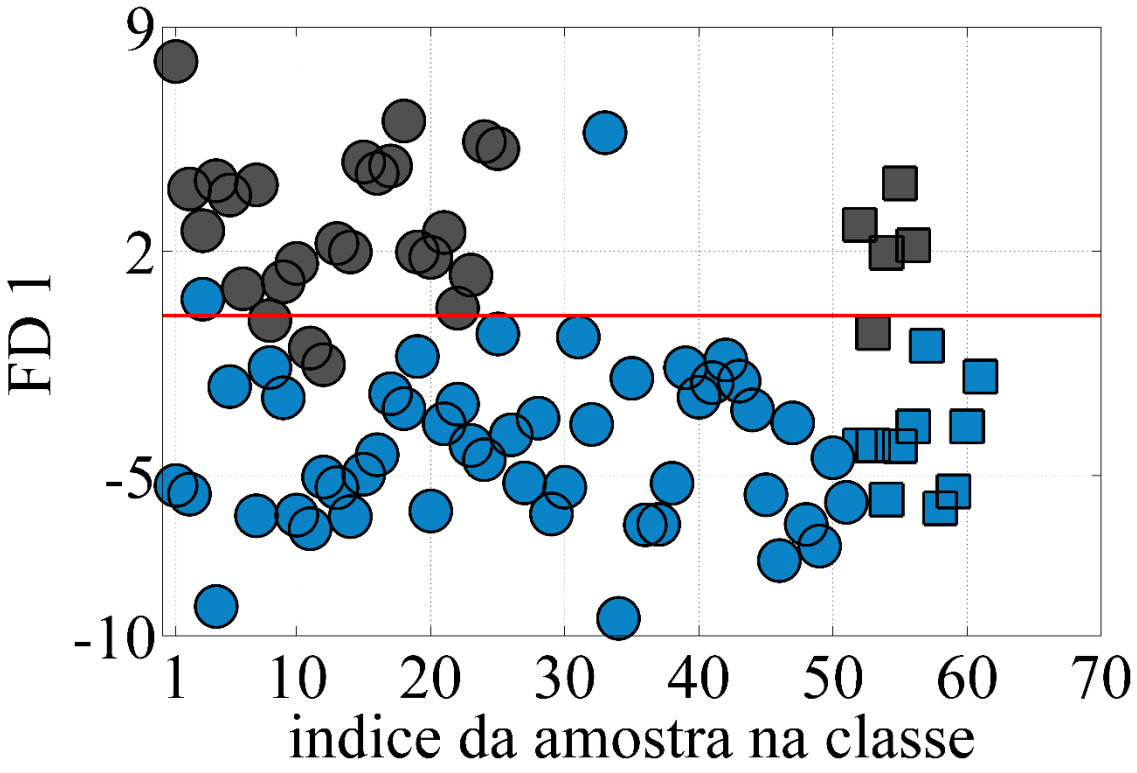

Figura 28 Os gráficos de funções discriminantes de Fisher obtidos para PCA-SW-LDA usando os espectros NIR das amostras não adulteradas (em azul) e adulteradas com extratos de madeira (em cinza). A linha vermelha indica o limite entre as classes, enquanto círculos e quadrados

representam as amostras nos conjuntos de treinamento e teste, respectivamente. amostras não adulteradas (em azul) e adulteradas com extratos de madeira (em cinza).

O modelo PCA-GA-LDA (**Figura 29**), apresentou valores de taxa de classificação correta de 97,4% para o conjunto de treinamento. Sete componentes principais foram selecionadas dentre as quais se encontram as PCs 3, 11 e 15, que possuem os três maiores valores de poder discriminante e são responsáveis por um melhor desempenho do modelo no conjunto de treinamento em comparação com o modelo PCA-SW-LDA que não incluiu a PC 15.

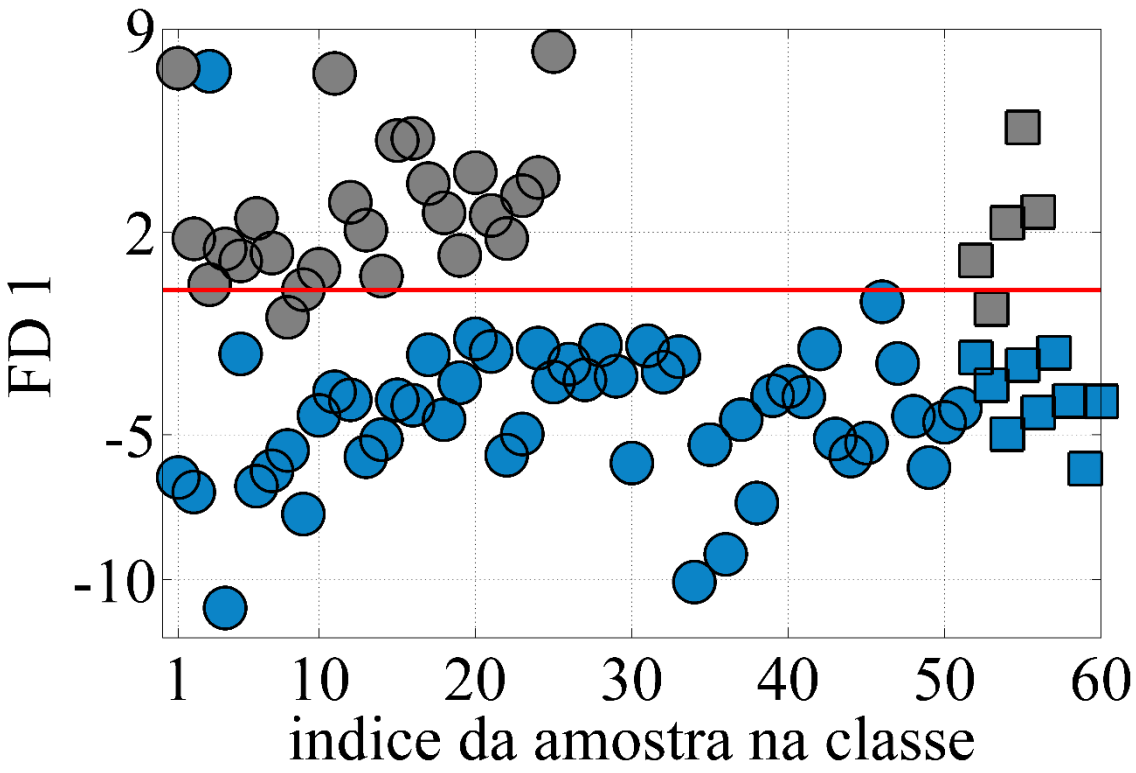

Figura 29 Os gráficos de funções discriminantes de Fisher obtidos para PCA-GA-LDA usando os espectros NIR das amostras não adulteradas (em azul) e adulteradas com extratos de madeira (em cinza). A linha vermelha indica o limite entre as classes, enquanto círculos e quadrados representam as amostras nos conjuntos de treinamento e teste, respectivamente. amostras não adulteradas (em azul) e adulteradas com extratos de madeira (em cinza).

O modelo PCA-GA-LDA incluiu ainda as PCs 8, 12, 16, 18 que de acordo com a **Figura 26b** não tem altos valores de discriminabilidade, que pode estar relacionado a menores valores

de taxa de classificação correta quando comparado ao modelo PCA-DISP-LDA que atingiu 100% de classificação correta. Na **Figura 30** é mostrado que o modelo não apresentou nenhum erro de classificação.

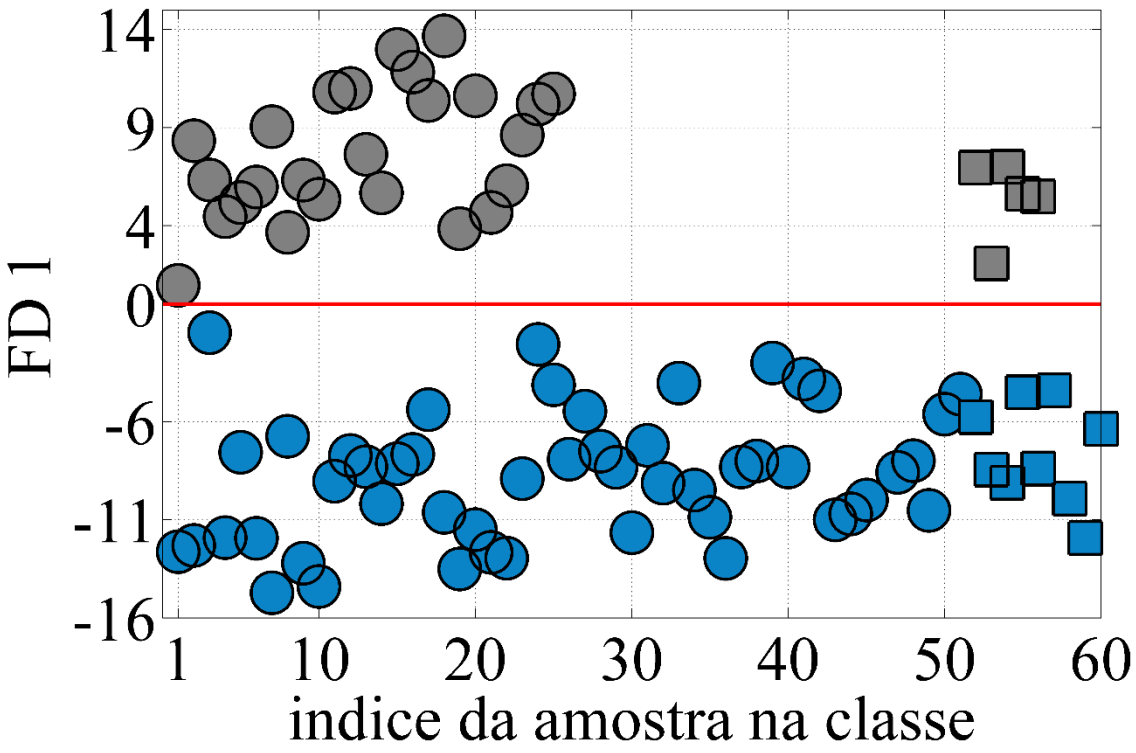

Figura 30 Os gráficos de funções discriminantes de Fisher obtidos para PCA-DISP-LDA usando os espectros NIR das amostras não adulteradas (em azul) e adulteradas com extratos de madeira (em cinza). A linha vermelha indica o limite entre as classes, enquanto círculos e quadrados representam as amostras nos conjuntos de treinamento e teste, respectivamente. amostras não adulteradas (em azul) e adulteradas com extratos de madeira (em cinza).

Para o modelo PCA-DISP-LDA foram utilizadas oito componentes principais, sendo elas 3, 4, 5, 6, 11, 15, 17 e 20. De acordo com o que é apresentado na **Figura 26b** todas as componentes principais de maior poder discriminante foram utilizadas no modelo e nenhuma componente principal de baixa discriminabilidade foi adicionada, dessa forma, nenhuma amostra adulterada ou não adulterada foi classificada de forma errada, o que atribui ao modelo PCA-DIP-LDA 100% de classificação correta entre todos os modelos testados neste estudo de caso.

Em comparação com as técnicas de seleção de variáveis o modelo PCA-DISP-LDA foi superior em termos de desempenho de classificação correta, embora tenha utilizado mais componentes que elas. A explicação para isso é que as componentes utilizadas possuíam poder considerável discriminante, sendo também a explicação para o melhor desempenho. Ainda levando em conta os modelos que utilizaram seleção de componentes, o modelo DISP foi capaz de deixar de fora as componentes que tinham baixo poder discriminante, gerando modelos com o número de componentes adequado.

De acordo com o gráfico apresentado na **Figura 30**, existe uma boa separação entre as duas classes de amostras, e pouquíssimas delas estão na região limite representada pela linha em vermelho, o que mostra que o modelo PCA-DISP-LDA está muito bem ajustado para este estudo de caso.

Capítulo 6

CONCLUSÃO

6. CONCLUSÃO

Neste trabalho foi proposto um critério de escolha de componentes principais, a fim de serem utilizados na modelagem por análise discriminante linear, como alternativa ao que é feito no modo convencional, em que as componentes principais são escolhidas em termos da variância explicada. Para tal, as componentes principais foram ordenadas de forma decrescente em termos de discriminabilidade de Fisher e então utilizadas como dados de entrada nos modelos LDA seguindo a ordem estabelecida nesta reordenação.

A estratégia foi avaliada frente a quatro estudos de caso, sendo um contendo dados simulados, um envolvendo a classificação de óleos de soja a respeito de estarem ou não em condições de consumo pela data de validade, um envolvendo a identificação da adição de óleo de soja em substituição ao biodiesel para ser utilizado como combustível e um último banco de dados referente a classificação de amostras de cachaça envelhecidas e não envelhecidas, adulteradas ou não com extratos de madeiras.

Para o primeiro conjunto de dados o método proposto foi assertivo em escolher a única componente principal que tinha informação discriminante, obtendo taxa de classificação de 100% para os conjuntos de treinamento e teste, além de apresentar um modelo livre de componentes que não contribuem efetivamente para melhorar a classificação.

Em relação ao estudo envolvendo a classificação de óleo de soja vencido e não vencido, o método proposto foi a técnica que apresentou a melhor relação entre taxa de classificação correta e parcimônia, errando apenas 1 amostra no conjunto de treinamento e utilizando duas componentes principais com melhor poder discriminante, quando comparado com as PCs utilizadas pelo método PCA-LDA, melhorando o desempenho nos conjuntos de treinamento e teste.

No terceiro banco de dados o método proposto assim como o método PCA-GA-LDA atingiu uma taxa de classificação correta de 100% para os conjuntos de treinamento e teste. O desempenho dos modelos foi consideravelmente superior ao método PCA-LDA, o que demonstra que a ordenação das PCs em termos de variância explicada contribui de forma menos efetiva para a classificação quando comparado ao método proposto.

Em comparação com o método PCA-GA-LDA o método proposto apesar de ter obtido o mesmo valor de classificação, atingiu o mesmo resultado com uma componente a menos, além de possuir uma simplicidade operacional maior. Uma vez que o método proposto não

requer nenhum tipo de ajuste de parâmetros e possui uma matemática muito mais simples, ganhando em esforço computacional para atingir o mesmo resultado.

Por último o estudo de caso envolvendo amostras de cachaça para classificação em termos da adulteração com extratos de madeira o método proposto foi o único modelo que atingiu a marca de 100% de amostras classificadas corretamente. Todas as componentes utilizadas tinham poder discriminante, diferente do que ocorreu com as demais técnicas.

É importante ressaltar que a superioridade do método proposto se dá em casos particulares em que a informação discriminativa não é responsável pela maior parte da variância explicada, e que caso contrário, ou seja, quando a informação discriminante for responsável majoritária pela variabilidade dos dados, o método proposto tende a se comportar de forma similar o método PCA-LDA, uma vez que as PCs de maior variância explicada serão também as mais discriminantes.

O método proposto neste trabalho também tem implementação mais simples quando comparado aos métodos de seleção de variáveis, uma vez que não requer a otimização de nenhum parâmetro de entrada, e pode ser adicionado diretamente no algoritmo PCA-LDA, já que requer apenas o cálculo da discriminabilidade e reordenação das PCs.

REFERÊNCIAS

REFERÊNCIAS

Allegretta, I., Marangoni, B., Manzaric, P., Porfidoa, C., Terzanoa, R., Pascaled, O., Senesid, G.S., (2020) Macro-classification of meteorites by portable energy dispersive X-ray fluorescence spectroscopy (pED-XRF), principal component analysis (PCA) and machine learning algorithms, Talanta 212-120785.
https://doi.org/10.1016/j.talanta.2020.120785

Almeida, V.E., Fernandes, D.D.S., Diniz, P.H.G.D., Gomes, A.D., Veras, G., Galvão, R.K.H., Araújo, M.C.U., (2021) Scores selection via Fisher's discriminant power in PCA-LDA to improve the classification of food data, Food Chem. 363-130296. https://doi.org/10.1016/j.foodchem.2021.130296

Ballabio, D., Consonni, V., (2013) Classification tools in chemistry. Part 1: linear models PLS-DA. Analitycal Methods. 5-3790-3798. https://doi.org/10.1039/C3AY40582F

Beuren, G.M., Anzanello, M.J., (2019) Variable selection using statistical non-parametric tests for classifying production batches into multiple classes, Chemometrics and Intelligent Laboratory Systems 193-103830.
https://doi.org/10.1016/j.chemolab.2019.103830

Caneca, A.R., Pimentel, M.F., Galvão, R.K.H., Matta, C.E., Carvalho F.R., Raimundo Jr, I.V., Pasquini, C., Rohwedder, J.J.R., (2006) Assessment of infrared spectroscscopy and multivariate techniques for monitoring the service condition of diesel-engine lubricating oils, Talanta 70-344 – 352.
https://doi.org/10.1016/j.talanta.2006.02.054

Chen, B-W., (2018) Incomplete data classification—Fisher discriminant ratios versus Welch discriminant ratios, Future Generation Computer Systems.
Chen, X., Hou, D., Zhao, L., Huang, P., Zhang, G., (2014) Study on defect classification in multi-layer structures based on Fisher linear discriminate analysis by using pulsed eddy current technique, NDT&E International 67-46 – 54.
https://doi.org/10.1016/j.ndteint.2014.07.003

Chi, J., Ma, Y., Weng, F.L., Thiessen-Philbrook, H., Parikh, C.R., Du, H., (2017) Surface-enhanced Raman scattering analysis of urine from deceased donors as a prognostic tool for kidney transplant outcome, J. Biophotonics 10-1743–1755.
https://doi.org/10.1002/jbio.201700019

Costa, G.B., Fernandes, D.D.S., Gomes, A.A., Almeida, V.E., Véras, G., (2016) Using near infrared spectroscopy to classify soybean oil according to expiration date. Food Chem. 196-539–543.
https://doi.org/10.1016/j.foodchem.2015.09.076

Duda, R.O., Hart, P.E., Stork, D.G., (2001) Pattern Classification, second ed., John Wiley, New York,. ISBN: 978-0-471-05669-0
https://www.wiley.com/en-ae/Pattern+Classification%2C+2nd+Edition-p-9781118586006

Fang, L., Zhao, H., Wang, P., Yu, M., Yan, J., Cheng, W., Chen, P., (2015) Feature selection method based on mutual information and class separability for dimension reduction in multidimensional time series for clinical data, Biomedical Signal Processing and Control 21-82 – 89.
https://doi.org/10.1016/j.bspc.2015.05.011

Fernandes, D.D.S., Almeida, V.E., Fontes, M.M., Araújo, M.C.U., Veras, G., Diniz, P.H.G.D. (2019) Simultaneous identification of the wood types in aged cachaças and their adulterations with wood extracts using digital images and SPA-LDA, Food Chem. 273-77–84.
https://doi.org/10.1016/j.foodchem.2018.02.035

Fernandes, D.D.S., Gomes, A.A., Fontes, M.M., Costa, G.B., Almeida, V.E., Araújo, M.C.U., Galvão, R.K.H., Véras, G., (2014) UV-Vis spectrometric detection of biodiesel/diesel blend adulterations with soybean oil, J. Braz. Chem. Soc. 25-169–175.
https://doi.org/10.5935/0103-5053.20130259

Foley, D.H., Sammon Jr, J.W., (1975) An optimal set of discriminant vectors, IEEE Transactions on Computers, 24-281 - 289.
https://doi.org/10.1109/T-C.1975.224208

Galvao, R.K.H., Araujo, M.C.U., José, G.E., Pontes, M.J.C., Silva, E.C., Saldanha. T.C.B., (2005)A method for calibration and validation subset partitioning, Talanta 67-736–740.
https://doi.org/10.1016/j.talanta.2005.03.025

Gao, Q., Liu, J., Zhang, H., Hou, J., Yang, X., (2012) Enhanced fisher discriminant criterion for image recognition, Pattern Recognition 45-3717 – 3724.
https://doi.org/10.1016/j.patcog.2012.03.024

Gomes, A.A., Galvão, R.K.H., Araújo, M.C.U., Veras, G, Silva, E.G., (2013) The Successive Projections Algorithm for invertval selection in PLS, Microchem. J, 110-202 -208.
https://doi.org/10.1016/j.microc.2013.03.015

Groger, T., Schaffer, M., Putz, M., Ahrens, B., Drew, K., Eschner, M., Zimmermann, R., (2008) Application of two-dimensional gas chromatography combined with pixel-based chemometric processing for the chemical profiling of illicit drug samples, Journal of Chromatography A, 1200-8 – 16.
https://doi.org/10.1016/j.chroma.2008.05.028

Hea, Z., Wub, M., Zhaoa, X., Zhang, S., Tan, J., (2021) Representative null space LDA for discriminative dimensionality reduction, Pattern Recognition 111-107664.
https://doi.org/10.1016/j.patcog.2020.107664

Huang, S., Yang, D., Yongxin, G., Zhang, X., (2015) Combined supervised information with PCAvia discriminative component selection, Information Processing Letters 115-812 – 816.
https://doi.org/10.1016/j.ipl.2015.06.010

Huang, Y., Liu, C., Zha, X.F., Li, Y., (2010) A lean model for performance assessment of machinery using second generation wavelet packet transform and Fisher criterion, Expert Systems with Applications 37-3815 – 3822.
https://doi.org/10.1016/j.eswa.2009.11.038

Kaur, T., Saini, B.S., Gupta, S., (2018) An optimal spectroscopic feature fusion strategy for MR brain tumor classification using Fisher Criteria and Parameter-Free BAToptimization algorithm, biocybernetics and biomedical engineering 38-409 – 424.
https://doi.org/10.1016/j.bbe.2018.02.008

Kaznowska, E., Depciuch, J., Łach, K., Kołodziej, M., Koziorowska, A., Vongsvivut, J., Zawlik, I., . Cholewa, M., Cebulski, J., (2018) The classification of lung cancers and their degree of malignancy by FTIR, PCA-LDA analysis, and a physics-based computational model, Talanta 186-337–345.
https://doi.org/10.1016/j.talanta.2018.04.083

Leardi, R., Norgaard, L., (2004) Sequuential Application of Backward Interval Partial Least Squares and Genetic Algorithms for the selection of relevant spectral regions, J. Chemometr. 18-4486 – 4497.
https://doi.org/10.1002/cem.893

Lee, L.C., Jemain, A.A., (2021) On overview of PCA application strategy in processing high dimensionality forensic data, Microchemical Journal 169-106608.
https://doi.org/10.1016/j.microc.2021.106608

Li, Y., Chen, S., Chen, H., Guo, P., Li, T., Xu, Q., (2020) Effect of thermal oxidation on detection of adulteration at low concentrations in extra virgin olive oil: Study based on laser-induced fluorescence spectroscopy combined with KPCA–LDA, Food Chemistry 309-125669.
https://doi.org/10.1016/j.foodchem.2019.125669

Li, C.Q., Fang, Z., Xu, Q.S., (2020) A partition-based variable selection in partial least squares regression, Chem. and Intel. Lab. Syst. 198-103935.
https://doi.org/10.1016/j.chemolab.2020.103935

Loog, M., (2007) On an alternative formulation of the Fisher criterion that overcomes the small sample problem, Pattern Recognition 40-1753 – 1755.
https://doi.org/10.1016/j.patcog.2006.11.002

Mahdianpari, M., Salehi, B., Mohammadimanesh, F., Brisco, B., Mahdavi, S., Amani, M., Granger, J. E, (2018) Fisher Linear Discriminant Analysis of coherency matrix for wetland classification using PolSAR imagery, Remote Sensing of Environment 206-300 – 317.
https://doi.org/10.1016/j.rse.2017.11.005

Manfredi, M., Robotti, E., Quasso, F., Mazzucco, E., Calabrese, G., Marengoa, E., (2018) Fast classification of hazelnut cultivars through portable infrared spectroscopy and chemometrics, Spec. Acta Part A 189-427-435.
https://doi.org/10.1016/j.saa.2017.08.050

Maugis, C., Celeux, G., Martin-Magniette, M.-L (2011) Variable selection in model-based discriminant analysis, J. Multivar. Anal 102, 1374–1387.
https://doi.org/10.1016/j.jmva.2011.05.004

Morais, C.L.M., Lima, K.M.G., Martin F.L., (2019) Uncertainty estimation and misclassification probability for classification models based on discriminant analysis and support vector machines, Anal. Chim. Acta 1063-40–46.
https://doi.org/10.1016/j.aca.2018.09.022

Pasquini, C., (2018) Near infrared spectroscopy: A mature analytical technique with new perspectives - A review, Anal. Chim. Acta 1026- 8–36.
https://doi.org/10.1016/j.aca.2018.04.004

R.A. Fisher, (1936) The use of multiple measurements in taxonomic problems, Ann. Eugen. 7-178–188.
https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Reile, C.G., Rodríguez, M.S., Fernandes, D.D.S., Gomes, A.A., Diniz, P.H.G.D., DI Anibal, C.V., (2020) Qualitative and quantitative analysis based on digital images to determine the adulteration of ketchup samples with Sudan I dye, in proof Food Chemistry.
https://doi.org/10.1016/j.foodchem.2020.127101

Ren, G., Ning, J., Zhang, Z., (2021) Multi-variable selection strategy based on near-infrared spectra for the rapid description of dianhong black tea quality, Spec. Acta Part A 245-118918.
https://doi.org/10.1016/j.saa.2020.118918

Rozza, A., Lombardi, G., Casiraghi, E., Campadelli, P., (2012) Novel Fisher discriminant classifiers, Pattern Recognition 45-3725 – 3737.
https://doi.org/10.1016/j.patcog.2012.03.021

Sampaio, S. L., Barreira, J.C.M., Fernandes, A., Petropoulos, S.A., Alexopoulos, A., Sa.-Buelga, C., Ferreira, I.C.F.R., Barros, L., (2021) Potato biodiversity: A linear discriminant analysis on the nutritional and physicochemical composition of fifty genotypes, Food Chemistry 345-128853.
https://doi.org/10.1016/j.foodchem.2020.128853

Sem, V., (2021) Interpretability of selected variables and performance comparison of variable selection methods in a polyethylene and polypropylene NIR classification task, Spectrochimica Acta Part A: 258-119850.
https://doi.org/10.1016/j.saa.2021.119850

Siqueira, L.F.S., Araújo Júnior, R.F., Araújo, A.A., Morais, C.LM., Lima, K.M.G. (2017) LDA vs. QDA for FT-MIR prostate cancer tissue classification, Chemometr. Intell. Lab.Syst. 162-123–129.
https://doi.org/10.1016/j.chemolab.2017.01.021

Tang, H., Fang, T., Shi P.-F., (2006) Laplacian linear discriminant analysis, Pattern Recognition 39-136–139.
https://doi.org/10.1016/j.patcog.2005.06.016

Tominaga, Y., (1999) Comparative of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and *k*-NN, Chemometr. Intell. Lab. Syst. 49-105–115.
https://doi.org/10.1016/S0169-7439(99)00034-9

Wang, Z., Hu, H., Zhang, L., Xue, J-H, (2018) Novel Fisher discriminant filtering (DGF) for hyperspectral image classification, Neurocomputing 275-1981 – 1987.
https://doi.org/10.1016/j.neucom.2017.10.046

Yang, J., Yang, J-Y., Zhang, D., (2002) What's wrong with Fisher criterion?, Pattern Recognition 35-2665 – 2668.
https://doi.org/10.1016/S0031-3203(02)00071-7

Ye, O., Rodionova, Pomerantsev, A.L., (2020) Chemometric tools for food fraud detection: The role of target class in nontargeted analysis, Food Chemistry 317-126448.
https://doi.org/10.1016/j.foodchem.2020.126448

Yun, Y.-H., Li, H.-D., Deng, B.-C., Cao, D.-S., (2019), An overview of variable selection methods in multivariate analysis of near-infrared spectra, Trends Anal. Chem. 113-102–115.
https://doi.org/10.1016/j.trac.2019.01.018

Zeng, X., Naghedolfeizi, M., Arora, S., Yousif, N., Aberra, D., (2016) Selection of principal components based on Fisher discriminant ratio, *Proc. SPIE 9871, Sensing and Analysis Technologies for Biomedical and Cognitive Applications* 98710K.
https://doi.org/10.1117/12.2227045

Zhang, P., Xu, Z., Wang, Q., Fan, S., Cheng, W., H. Wang, Y. W, (2021) A novel variable selection method based on combined moving window and intelligent optimization algorithm for variable selection in chemical modeling, Spectr. Acta Part A, 246-118986.
https://doi.org/10.1016/j.saa.2020.118986

Zheng, W.-S., Lai. J.-H., Yuen, P.C., (2005) GA-Fisher: a new LDA-based face recognition algorithm with selection of principal components, IEEE T. Syst. Man Cy. B 35-1065–1078.
https://doi.org/10.1109/TSMCB.2005.850175.

APÊNDICE

APÊNDICE 1: SCORES E LOADINGS REFERENTES AOS BANCOS DE DADOS UTILIZADOS

DADOS SIMULADOS

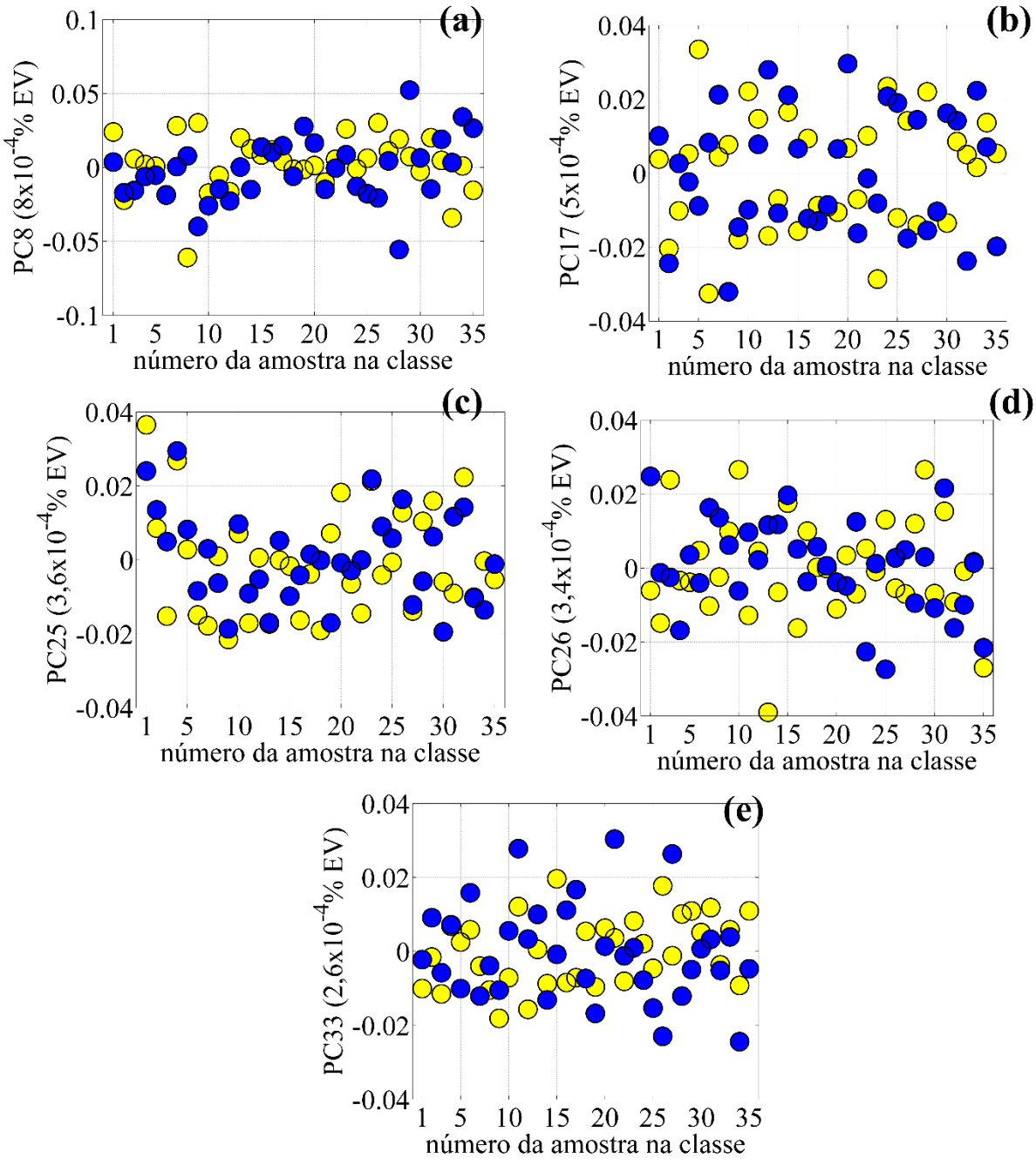

Figura 31 Dados simulados: (a-e) Scores de PCA para as componentes principais (PCs) 8,17,25,26 e 33 ordenadas de forma crescente de variância explicada.

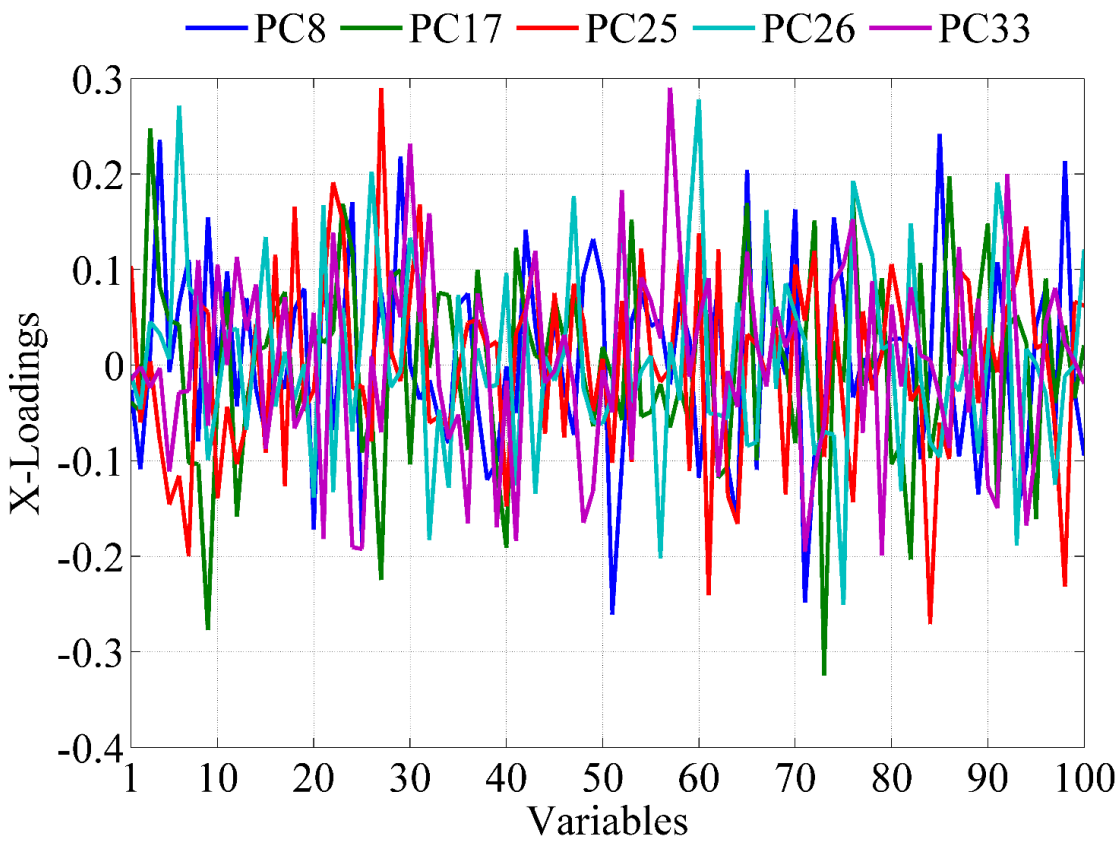

Figura 32 Dados simulados: loadings de PCA para as PCs 8, 17, 25,26 e 33.

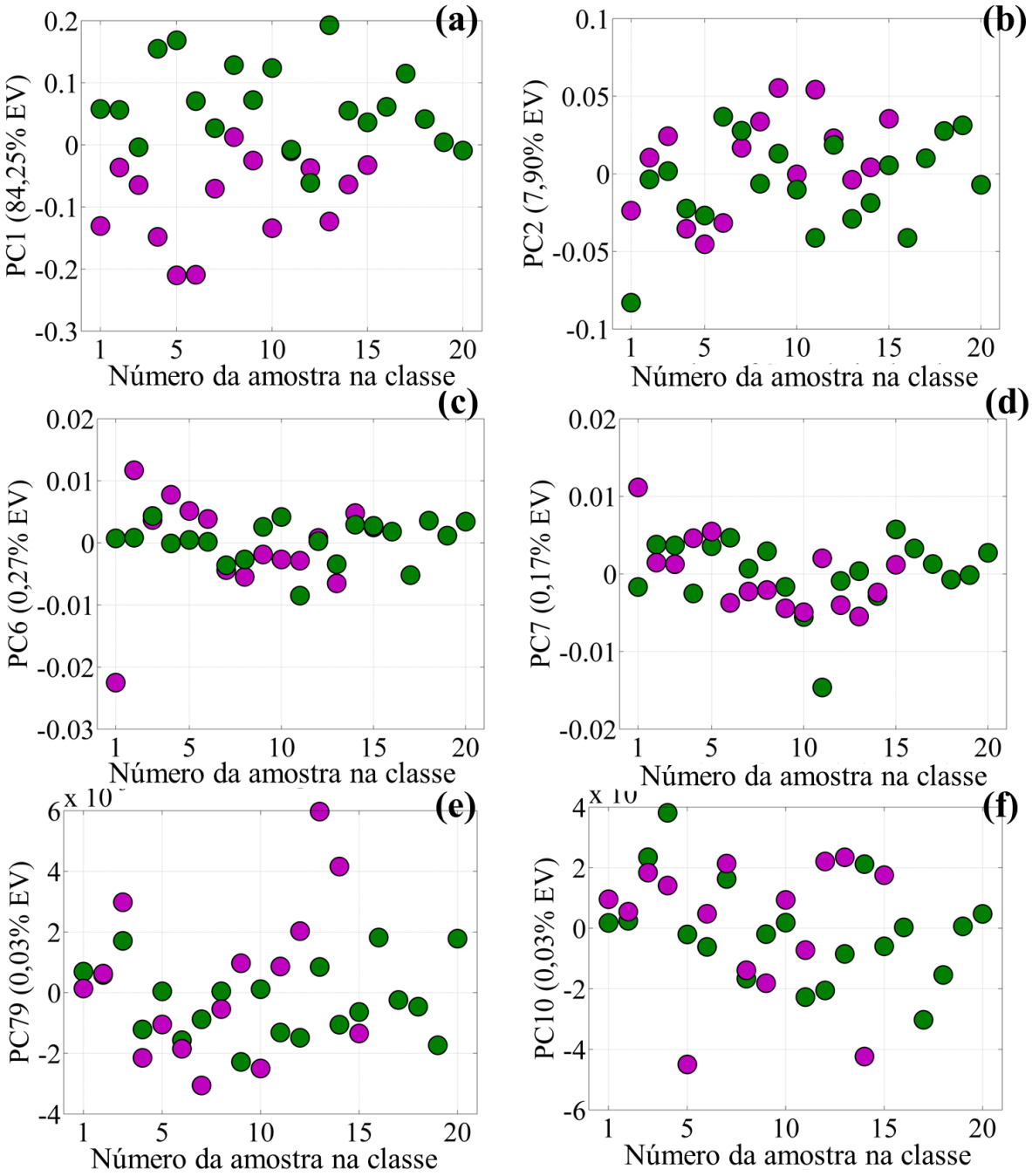

Figura 33 Classificação de óleos de soja em relação à data de validade por espectroscopia NIR: (a-*f*) Scores de PCA para as componentes principais (PCs) 1,2,6,7, 9 e 10 ordenadas de forma crescente de variância explicada.

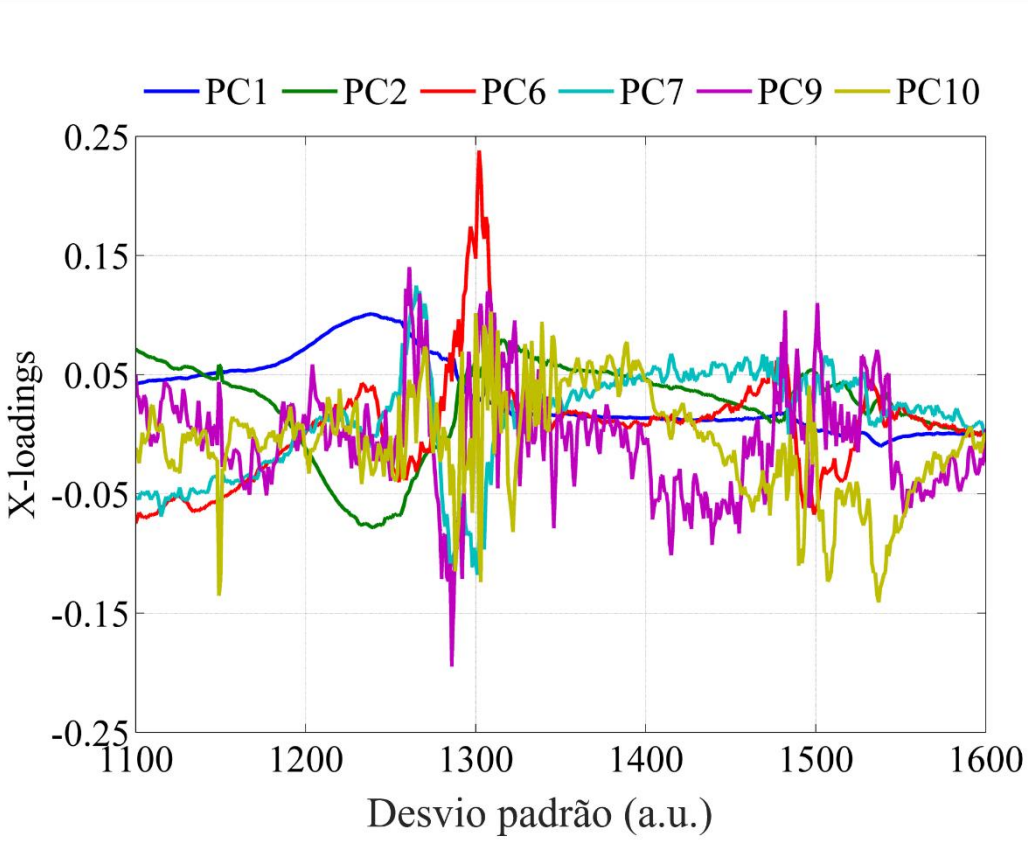

Figura 34 Classificação de óleos de soja em relação à data de validade por espectroscopia NIR: loadings de PCA para as PCs 1, 2, 6,7,9 e 10.

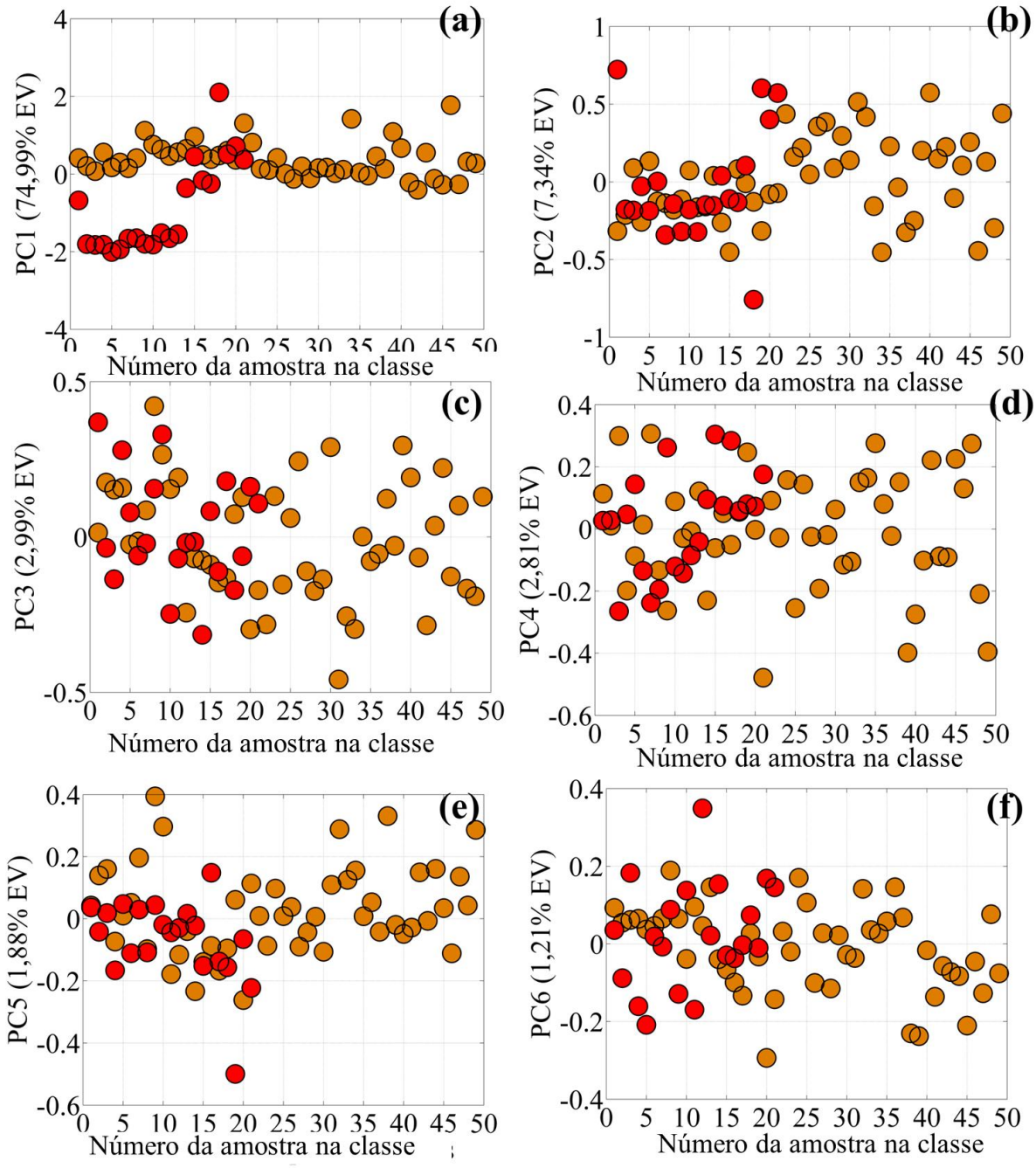

Figura 35 Identificação da adulteração de misturas de biodiesel/diesel com óleo de soja por espectroscopia Vis-NIR: (a-*f*) Scores de PCA para as componentes principais (PCs) 1 a 6 ordenadas de forma crescente de variância explicada.

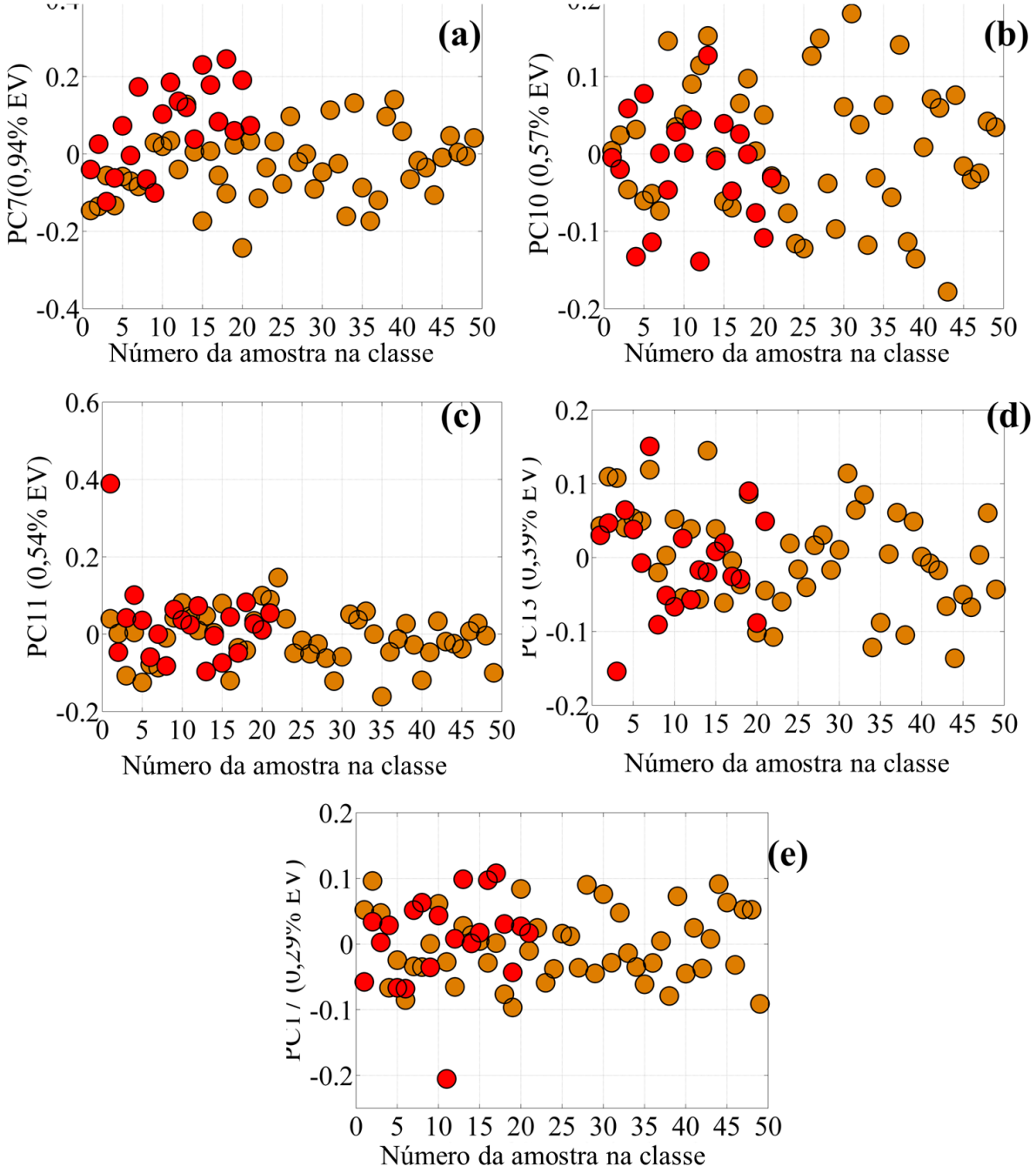

Figura 36 Identificação da adulteração de misturas de biodiesel/diesel com óleo de soja por espectroscopia Vis-NIR: (a-*f*) Scores de PCA para as componentes principais (PCs) 7,10,11,13 e 17 ordenadas de forma crescente de variância explicada.

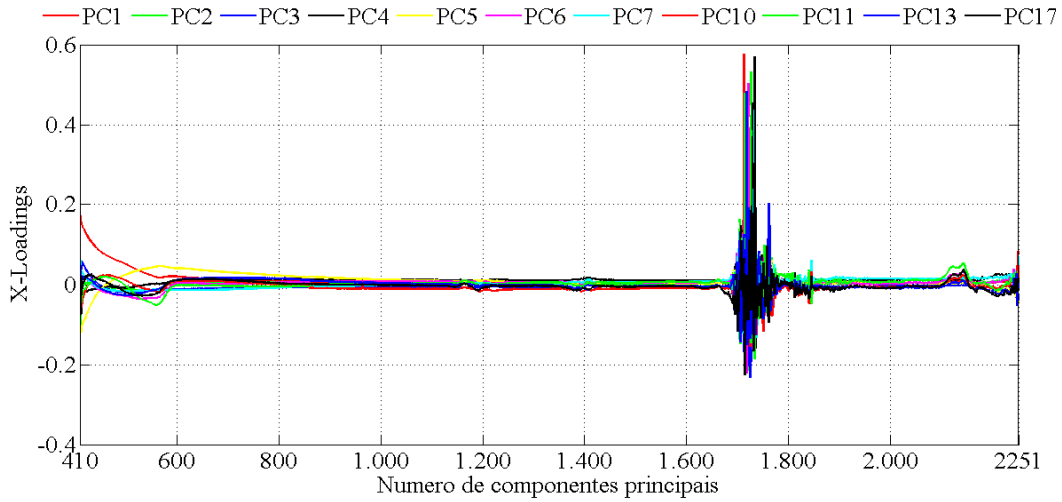

Figura 37 Identificação da adulteração de misturas de biodiesel/diesel com óleo de soja por espectroscopia Vis-NIR: loadings de PCA para as PCs 1 a 7, 10, 11, 13 e 17.

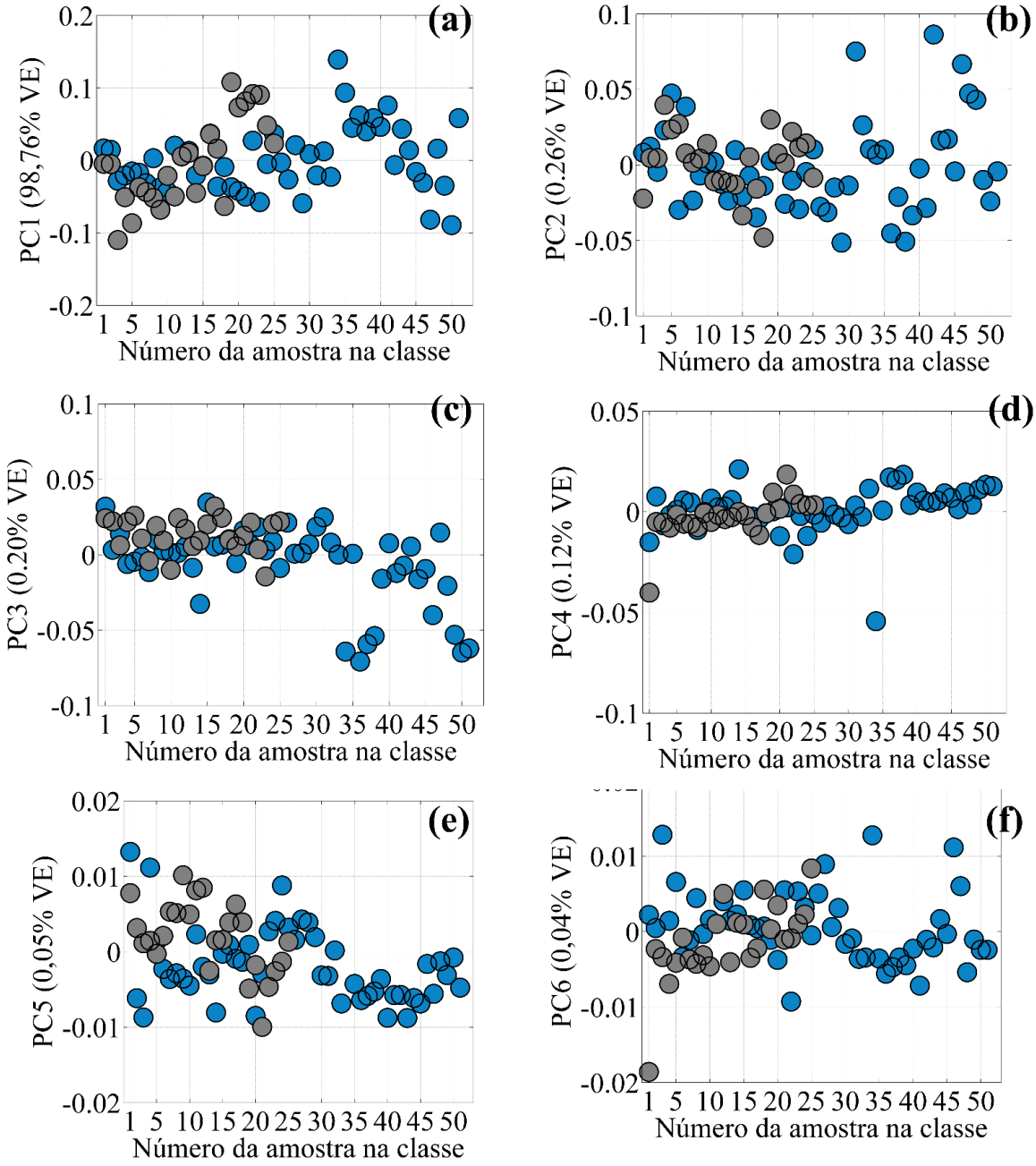

Figura 38 Identificação da adulteração de cachaça quanto ao tempo de envelhecimento em barrís de madeira: (a-f) Scores de PCA para as componentes principais (PCs) 1 a 6 ordenadas de forma crescente de variância explicada.

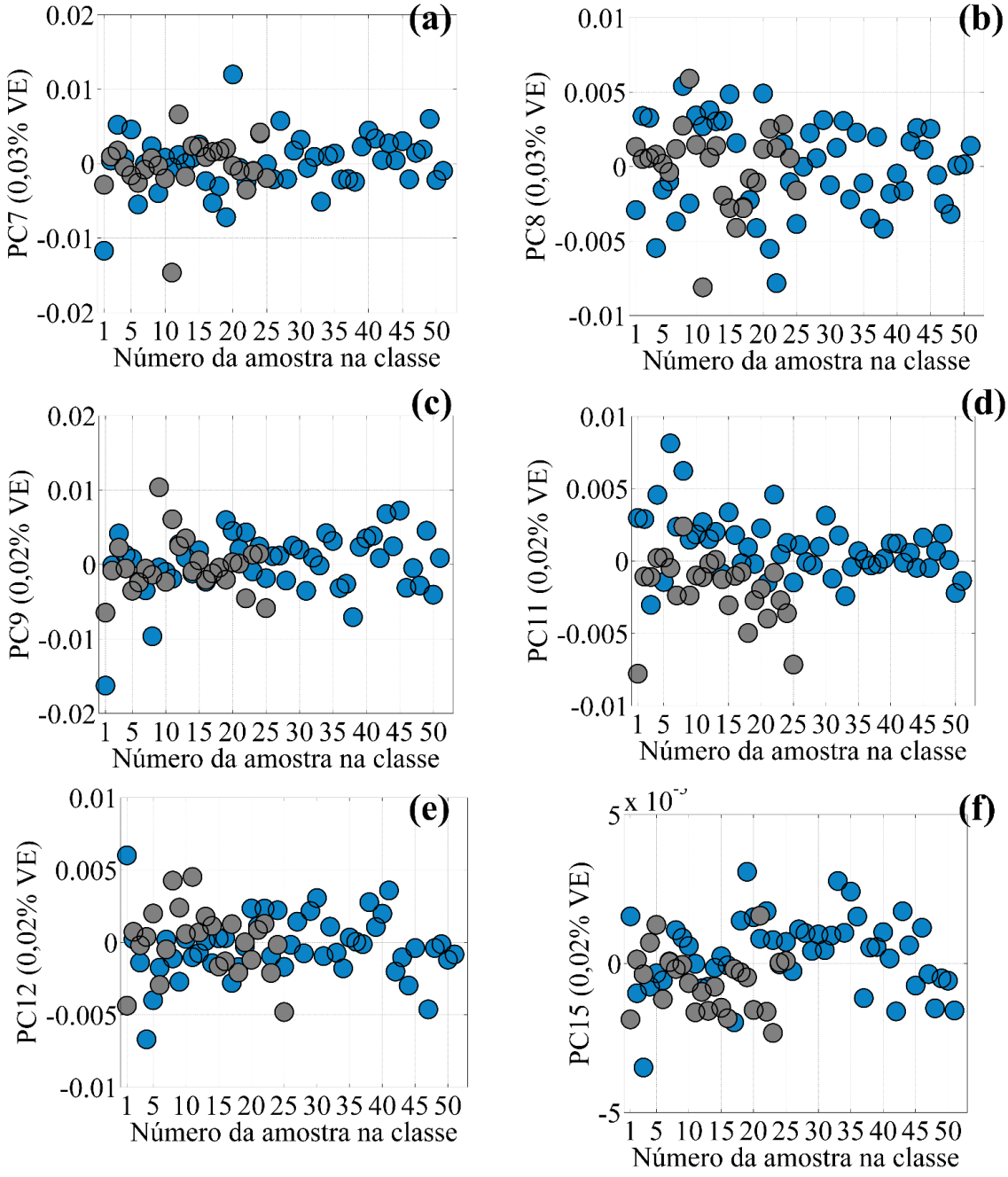

Figura 39 Identificação da adulteração de cachaça quanto ao tempo de envelhecimento em barrís de madeira: (a-f) Scores de PCA para as componentes principais (PCs) 7 – 9, 11,12 e 15 ordenadas de forma crescente de variância explicada.

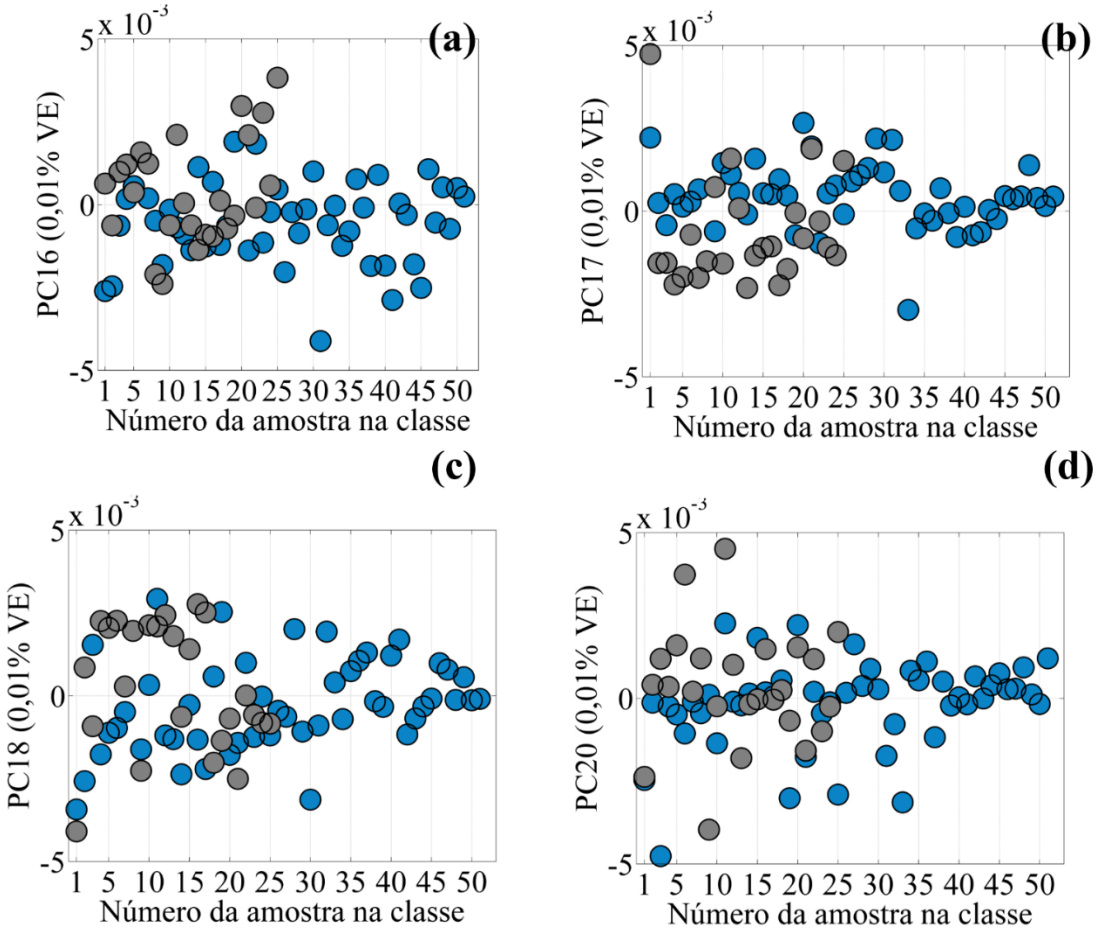

Figura 40 Identificação da adulteração de cachaça quanto ao tempo de envelhecimento em barrís de madeira: (a-d) Scores de PCA para as componentes principais (PCs) 16, 17, 18 e 20 ordenadas de forma crescente de variância explicada.

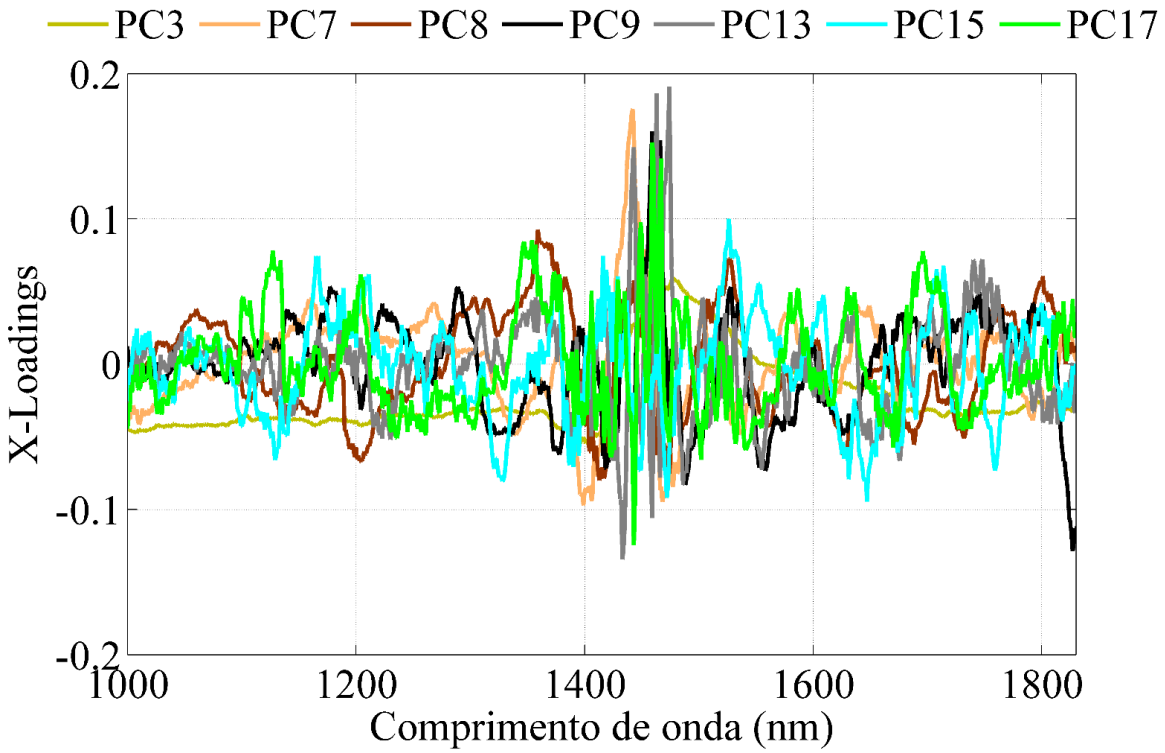

Figura 41 Identificação da adulteração de cachaça quanto ao tempo de envelhecimento em barrís de madeira: loadings de PCA para as PCs 3, 7, 8, 9, 13, 15 e 17.