UNIVERSIDADE FEDERAL DE PERNAMBUCO

Centro de Ciências Exatas e da Natureza – CCEN

Departamento de Química Fundamental – DQF

JOSÉ FRANCIELSON QUEIROZ PEREIRA

ANÁLISE DE ALTERAÇÕES DOCUMENTAIS USANDO IMAGENS HIPERESPECTRAIS NA REGIÃO DO INFRAVERMELHO MÉDIO E PRÓXIMO

Recife

2015

JOSÉ FRANCIELSON QUEIROZ PEREIRA

ANÁLISE DE ALTERAÇÕES DOCUMENTAIS USANDO IMAGENS HIPERESPECTRAIS NA REGIÃO DO INFRAVERMELHO MÉDIO E PRÓXIMO

Dissertação de Mestrado submetida ao programa de Pós-graduação em Química da Universidade Federal de Pernambuco, como parte dos requisitos para obtenção do título de Mestre em Química.

Orientadora: Maria Fernanda Pimentel
Co-orientador: Ricardo Saldanha Honorato

Recife

2015

Catalogação na fonte
Bibliotecária Elaine Cristina de Freitas CRB4-1790

P436a Pereira, José Francielson Queiroz
Análise de alterações documentais usando imagens hiperespectrais na região do infravermelho médio e próximo / José Francielson Queiroz Pereira . – 2015.
99 f.: fig., tab.

Orientadora: Maria Fernanda Pimentel Avelar
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN. Química Fundamental. Recife, 2015.
Inclui referências e Apêndices.

1. Química Analítica. 2. Imagem hiperespectral 3. Infravermelho. Forense . 4. Documentoscopia I. Avelar, Maria Fernanda Pimentel (Orientadora). II. Título.

543 CDD (22. ed.) UFPE-FQ 2018-15

JOSÉ FRANCIELSON QUEIROZ PEREIRA

ANÁLISE DE DOCUMENTOS QUESTIONÁVEIS USANDO IMAGENS HIPERESPECTRAIS NA REGIÃO DO INFRAVERMELHO MÉDIO E PRÓXIMO

Dissertação de Mestrado submetida ao programa de Pós-graduação em Química da Universidade Federal de Pernambuco, como parte dos requisitos para obtenção do título de Mestre em Química.

Aprovada em 28/08/2015

BANCA EXAMINADORA

_____

Profª Maria Fernanda Pimentel Avelar (Orientadora)
Departamento de Química Fundamental
Universidade Federal de Pernambuco

_____

Profª Elaine Cristina Lima do Nascimento
Unidade Acadêmica de Serra Talhada
Universidade Federal Rural de Pernambuco

_____

Drª Flávia Sousa de Lins Borba
Departamento de Química Fundamental
Universidade Federal de Pernambuco

AGRADECIMENTOS

Agradeço primeiramente a minha orientadora, Profª Drª Maria Fernanda Pimentel, que desde o primeiro contato transmitiu confiança e comprometimento com minha formação. Em pouco menos de dois anos de orientação ampliou meu horizonte para novas perspectivas acadêmicas e pessoais, além de contribuir enormemente para minha consolidação enquanto pesquisador. Sou muito grato por seus valorosos ensinamentos e sua dedicação.

Ao meu co-orientador o perito Dr. Ricardo Saldanha Honorato, meus sinceros agradecimentos pelas ideias e conselhos durante as discussões. Sempre levantou questões importantes e apresentou soluções valiosas para a conclusão desta pesquisa.

Ao prof. Dr. Célio Pasquini pelo apoio e encorajamento durante algumas etapas.

A CNPq, pela bolsa concedida, ao INCTAA e ao NUQAAPE, pelo apoio ao projeto, ao Grupo de Instrumentação e Automação (GIA), pelo apoio durante as análises e a UFPE/DQF pelo suporte institucional.

Ao Laboratório de Combustíveis (LAC) e toda sua equipe que possibilitaram a realização da pesquisa.

Aos amigos do DQF que me proporcionaram muitos e bons momentos de discussão, estudo e descontração. Arthur, Eivisson, Crislaine, Imerson, Leandro, Leonardo e Yaicel que são amigos para uma vida inteira.

Aos amigos e colegas de trabalho Alianda, Allan, Carolina, Eduardo, Jéssica, Leandro, Lívia, Marcela, Neirivaldo, Rafaella e Vitor por todos os conselhos, ensinamentos e discussões que me ajudaram a vencer vários desafios. Minha pesquisa é, sem dúvidas, um produto desse ambiente acolhedor que me foi oferecido por eles. Sou eternamente grato por todo o apoio que recebi no momento em que mais precisei.

Um agradecimento especial a Carolina por ter se prontificado a me ajudar em inúmeras situações, pelo treinamento técnico dado e por sua disponibilidade em tirar minhas dúvidas.

A minha amiga Alianda por ter realizado medidas essenciais para a conclusão deste trabalho.

Aos colegas do LAC, Carol, João, Vanessa, Gisele e Kaline que me ajudaram a resolver incontáveis questões de trabalho e sempre estiveram dispostos a ajudar.

As minhas amigas Paula, Polyana, Camila, Fabrícia e Ismália pelos favores prestados e bons momentos de descontração.

Aos meus companheiros de apartamento e grandes amigos, irmãos adotivos, Cícero e Geane que me acolheram e apoiaram desde o primeiro dia desta jornada.

A Andréa pela amizade, companheirismo, paciência e apoio nos momentos difíceis.

A profª Andréa por seu apoio e incentivo, pelas caronas e bons conselhos que me trouxeram até aqui.

Um agradecimento especial a Patrícia por sua dedicação, profissionalismo e por ter me apresentado o DQF, sou grato por ter sido a primeira pessoa que conheci no departamento.

A todos da minha família pelo incentivo em continuar trilhando o caminho acadêmico e buscando as minhas realizações pessoais e profissionais.

A minha mãe, por ser minha maior incentivadora e fonte de determinação, e ao meu pai que se orgulhava de ter me dado à educação que nunca.

RESUMO

A análise de documentos adulterados é um problema recorrente em departamentos de polícia científica. Os textos produzidos com canetas são foco de muitas investigações que buscam identificar alterações feitas por adição ou subtração de texto. É importante que a análise destas alterações preserve a integridade física das provas. Nesse sentido, a utilização de imagens hiperespectrais na região do infravermelho médio (HSI-MIR) e próximo (HSI-NIR) associadas a ferramentas quimiométricas de reconhecimento de padrão não supervisionadas se apresenta como uma estratégia não destrutiva capaz de extrair importantes informações sobre a composição química e a distribuição espacial dos compostos na amostra. Considerando que há uma dificuldade na discriminação de tintas de canetas pretas, a primeira etapa deste trabalho foi focada na avaliação da discriminação entre diferentes canetas com tinta de cor preta. A segunda etapa constou na realização de testes-cego para validar as metodologias de discriminação. Foram usadas 16 canetas de tinta preta com diferentes diâmetros de pontas, diferentes marcas e modelos para produzir linhas em papel sulfite branco. Em seguida, o acessório de Refletância Total Atenuada (ATR) foi utilizado para adquirir imagens hiperespectrais das linhas produzidas na faixa de 4000-750 cm^{-1}. O conjunto de dados foi pré-processado e, em seguida, combinados dois a dois, gerando 120 combinações. As técnicas não supervisionadas de análise exploratória "Análise de componentes principais (PCA)" e "Projection pursuit" (PP) foram aplicadas como ferramentas de reconhecimento de padrões em cada par, buscando discriminá-los. Na segunda etapa da investigação, as mesmas 16 canetas foram utilizadas por colaboradores para escrever números de até cinco dígitos em uma folha de cheque e em uma folha de papel branco. Os colaboradores utilizaram uma ou duas canetas para produzir cada número, simulando amostras não adulteradas e falsificadas, respectivamente. Com base na observação de cada número, foram adquiridas Imagens Hiperespectrais no Infravermelho Médio (HSI-MIR), em diferentes posições dos números. Depois de aplicar diferentes pré-processamentos e realizar a seleção de pixels, os conjuntos de dados foram submetidos à PCA e PP. Também foram adquiridas Imagens Hiperespectrais na faixa de 928-2524 nm. As imagens hiperespectrais no NIR foram pré-processadas com Variável Normal Padrão (SNV) e depois foram submetidas à PCA e análise de PP. Na primeira etapa, foi possível discriminar 89 combinações usando PCA e 117 usando PP, assim esta ultima técnica provou ser mais robusta e eficiente. Na segunda etapa da pesquisa, PCA associada à HSI-MIR e HSI-NIR foi possível solucionar 50,0% e 76,7% dos testes-cego, respectivamente. Quando PP foi associada à HSI-MIR e HSI-NIR os percentuais de acertos dos testes subiram para 63,3% e 83,3% respectivamente. O número de acertos usando HSI-NIR foi superior, no entanto, em alguns casos apenas com HSI-MIR foi possível solucionar o teste-cego. Combinando os resultados obtidos com HSI-MIR e HSI-NIR associadas com PP, o percentual de acerto chegou a 90% dos casos, demonstrando o potencial da metodologia para a análise de alterações documentais.

Palavras-chave: Imagem hiperespectral. Infravermelho. Forense. Documentoscopia. Análise multivariada.

ABSTRACT

The analysis of forged document is a common problem in scientific police departments. Texts produced with pens are the focus of many investigations that aiming to identify adulterations made by adding or subtracting text. It is important that analysis of these adulterations preserve the physical integrity of evidences. Therefore, hyperspectral imaging at middle infrared (HSI-MIR) and near infrared (HSI-NIR) associated with unsupervised pattern recognition methods represent a non-destructive strategy, able to extract important information about composition and spatial distribution of compounds in the sample. Considering the difficulty in discriminating among black ink pens, the first step of this work was focused on this purpose. The second step consisted in blind tests performances to validate the discrimination methodology. Sixteen black ink pens with different tip diameters, of different makes and models, were used to draw lines on white paper. Then, an ATR accessory was used to acquire hyperspectral images in the range of 4000 – 750 cm^{-1} from the line drawn. The datasets were preprocessed and combined two by two, producing 120 combinations. The unsupervised exploratory data analysis techniques "Principal component analysis (PCA)" and "Projection Pursuit (PP)" were applied to each pair aiming at cluster identification. In the second investigation step, the same 16 pens were used by collaborators to write numbers up to five digits on bank checks and sheets of white paper. The collaborators used one or two pens to produce each number, simulating unadulterated and forged samples, respectively. Based on observations of each number, HSI-MIR were acquired in different positions of the numbers. After apply different preprocess techniques and pixel selection into the dataset, PCA and PP were performed. HSI-NIR at 928-2524 nm range were also acquired. The hyperspectral images were preprocessed with standard normal variate (SNV) and PCA and PP were performed. The first step resulted in the discrimination of 89 combinations using PCA and 117 using PP, thus the latter technique proved to be more robust and efficient. In the second step of the research, PCA associated to HSI-MIR and HSI-MIR was able to solve 50.0% and 76.7% of the blind tests, respectively. When PP was associated to HSI-MIR and HSI-NIR the successfully identification percentage of the tests increased to 63.3% and 83.3% respectively. The number of correct identifications using HSI-NIR was higher. In some cases, however, only HSI-MIR was able to solve the blind test. Combining the results obtained with HSI-MIR and HSI-NIR associated with PP the percentage of correct identification reached 90% of the cases, demonstrating the potential of the methodology for document modification analysis.

Key-words: Hyperspectral imaging. Infrared. Forensic. Documentoscopy. Multivariate Analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ALS	Mínimos Quadrados Alternados (*Alternating Least Squares*)
ATR	Reflectância Total Atenuada (*Atenuate Total Reflectance*)
DA	Análise discriminante (*Discriminant Analysis*)
FIR	Infravermelho distante (*Far Infrared*)
FT-IR	Infravermelho com Transformada de Fourier (*Fourier Transformed –infrared*)
HSI	Imagem Hiperespectral (*Hyperespectral Imaging*)
LDA	Análise Discriminante Linear (Linear Discriminant Análysis)
MCR	Resolução de Curvas Multivariadas (*Multivariate Curve Resolution*)
MIR	Infravermelho Médio (*Middle Infrared*)
MSC	Correção de Espelhamento Multiplicativo (*Multiplicative Scatter Correction*)
NIR	Infravermelho Próximo (*Near Infrared*)
PA	Análise de Procusto (*Procrustes Analysis*)
PC	Componente Principal (*Principal Componet*)
PCA	Análise de Componentes Principais (*Principal Component Análysis*)
PLS-DA	Análise Discriminante por Mínimos Quadrados Parciais (*Partial Least Square Discriminant Analysis*)
PP	Projection Pursui
RCC	Razão de Classificação Correta
RGB	Vermelho, Verde e Azul (*Red, Green and Blue*)
ROI	Região de Interesse (*Region of Interst*)
SG	Derivada de Savitzky Golay (*Savitzky Golay Derivative*)
SNV	Variação Normal Padrão (*Standart Normal Variate*)
UV-vis	Regiões espectrais do Ultravioleta e do Visível
VP	Vetor de Projeção
VSC	Vídeo Comparador Espectral (*Video Spectral Comparator*)

SUMÁRIO

1 INTRODUÇÃO

Uma quantidade considerável de processos investigativos se baseia na análise e interpretação de provas documentais, sejam eles documentos oficiais, textos ou apenas fragmentos de texto com importância pericial. Contratos, procurações, bilhetes e cheques são alguns exemplos de documentos que podem ser usadas como evidência de uma fraude ou crime (BRAZ et al., 2013), para tanto a autenticidade desses documentos deve ser atestada. A análise e verificação de adulterações em documentos faz parte de um ramo da criminalística denominado Documentoscopia (BRUNELLE; CRAWFORD, 2003).

A falsificação de documentos manuscritos por meio da adição de texto ou modificações no texto original é uma modalidade de fraude documental bastante comum. Em geral esse tipo de fraude é realizada utilizando-se uma caneta com tinta visualmente similar a utilizada no texto original, para evitar a detecção da adulteração por simples inspeção visual. No entanto, as tintas de caneta são misturas muito complexas e quimicamente bastante diversificadas (SILVA et al., 2013). Assim a análise química da composição de tintas usadas para produzir documentos pode fornecer importantes informações sobre a autenticidade dos mesmos (EZCURRA et al., 2010).

Os métodos atuais para análise forense de tintas são a microscopia, fonte de luz alternada, a cromatografia líquida, cromatografia de camada delgada e espectroscopia de infravermelho (MURO et al., 2015). Embora bastante empregada na análise de tintas, a cromatografia de modo geral compromete a integridade das amostras, uma vez que necessita da remoção de parte da tinta do documento para ser realizada. Por esse motivo os métodos não destrutivos como as técnicas espectroscópicas são mais vantajosas.

Nam e colaboradores (2014) analisaram o perfil espectral de 63 canetas esferográficas de tinta preta contidas em um banco de dados e foram capazes de distinguir entre cada uma das canetas em função de marca, modelo e lote. Os autores buscaram em cada espectro sinais que possibilitassem a discriminação entre as tintas. Dessa forma a metodologia depende bastante da habilidade do avaliador e do tempo necessário para avaliar cada espectro.

Unidades de polícia científica utilizam vídeo-comparadores espectrais que empregam câmeras digitais, lâmpadas, espelhos e filtros na região do UV, visível e parte do infravermelho (SILVA et al., 2014a). Embora esse equipamento permita solucionar diversos casos, a metodologia da análise se baseia na inspeção visual ou seleção de poucos

comprimentos de onda para buscar diferenças entre as tintas. Dessa forma a análise é limitada por ser univariada e depender da habilidade do analista. Diante desta limitação, recentemente diferentes ferramentas de análise multivariada vêm sendo empregadas e aprimoradas constantemente para tornar as metodologias de análise mais rápidas e menos subjetivas.

A espectroscopia na região do infravermelho é uma técnica que vem se mostrando promissora na análise forense de tintas de caneta (MURO et al., 2015). Silva et al. (2013), utilizaram a espectroscopia no Infravermelho Médio (MIR, do inglês *Middle Infrared*) e Análise Discriminante Linear (LDA, do inglês *Linear Discriminant Analysis*) associadas a técnicas de seleção de variáveis para diferenciar tintas de caneta azul de acordo com o tipo e a marca em dois papeis diferentes. Os autores conseguiram separar 100% das canetas em função da marca. A correta classificação em relação ao tipo foi de 100% e 97,3% para os dois tipos de papel (SILVA et al., 2013).

Uma técnica de aquisição de espectros de infravermelho que vem se desenvolvendo e ganhando espaço entre as técnicas espectroscópicas é a técnica de Imagens Hiperespectrais (HSI, do inglês *Hyperspectral Imaging*). As Imagens Hiperespectrais na região do Infravermelho Próximo (HSI-NIR, do inglês *Hyperespectral Images in Near Infrared*) e do Infravermelho Médio (HSI-MIR, do inglês *Hyperespectral Images in Middle Infrared*) possibilitam obter uma grande quantidade de informação sobre a distribuição e composição da amostra e já possui aplicação em diferentes áreas. Na produção de fármacos estudos buscam determinar a distribuição dos compostos ativos e avaliar o processo de produção usando HSI (ELLISON et al., 2008; CRUZ et al., 2009; AMIGO; RAVN, 2009). Na química forense imagens hiperespectrais foram utilizadas para identificar fraudes documentais, determinar o tempo de degradação de sangue em cenas de crime e detecção de impressão digital (SILVA et al., 2014a; LI et al., 2013; TAHTOUH et al., 2007). Também vem sendo avaliada a possibilidade de utilização de HSI no controle de qualidade e segurança de diferentes alimentos (GOWEN et al., 2007). Essas e muitas outras aplicações evidenciam o potencial de HSI-NIR e HSI-MIR para resolução de problemas da documentoscopia.

As imagens hiperespectrais são conjuntos de dados que guardam grande quantidade de informação espacial e química sobre as amostras, mas para acessar as informações mais importantes é necessário utilizar diferentes ferramentas quimiométricas. Em se tratando especificamente de Documentoscopia é preferível que essas ferramentas sejam capazes de extrair as informações sem qualquer conhecimento prévio sobre as amostras. A Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*) é uma ferramenta de análise exploratória que pode ser usada como técnica de reconhecimento de padrão não supervisionada, que possibilita reduzir a dimensionalidade dos dados conservando as principais informações, permite identificar a correlação entre as variáveis, similaridade entre amostras, bem como avaliar e identificar as variáveis mais importantes para explicar a variância dos dados (BEEBE et al., 1998). Na PCA as amostras são projetadas em um novo sistema de eixos ortogonais entre si, cujas direções levam em conta os caminhos de projeção com máxima variância dos dados. Cada eixo/vetor de projeção é otimizado em função da máxima variância dos dados.

A análise de Project Pursuit (PP) é outra ferramenta de análise exploratória que pode ser utilizada como técnica de reconhecimento de padrão não supervisionada e que permite acessar diferentes informações sobre conjuntos de dados como as imagens hiperespectrais (HOU; WENTZELL, 2013). A análise de PP também cria um novo sistema de eixos para projetar as amostras, no entanto, diferente da PCA, a técnica de PP utiliza a *curtose* como critério para otimizar os vetores de projeção. Wentzell e colaboradores (2015) demonstraram que a análise de PP pode ser útil para discriminar canetas de tinta azul utilizando os espectros de infravermelho médio obtidos por refletância total atenuada.

2 OBJETIVO GERAL

Utilizar as imagens hiperespectrais na região do infravermelho próximo (HSI-NIR) e infravermelho médio (HSI-MIR), associadas a técnicas de reconhecimento de padrão não supervisionadas, para desenvolver uma metodologia rápida, confiável e não destrutiva para a análise de documentos.

2.1 Objetivos Específicos

Avaliar o potencial de HSI-MIR associada às técnicas de reconhecimento de padrão não supervisionadas PCA e PP para discriminação de canetas distintas com tinta de cor preta.

Avaliar o uso das técnicas HSI-NIR e HSI-MIR associadas á ferramentas de reconhecimento de padrão PCA e PP para identificação de falsificações em amostras de teste-cego (com amostras de composição desconhecida pelo analista).

3 FUNDAMENTAÇÃO

3.1 Infravermelho

A espectroscopia no infravermelho (IR) é uma importante ferramenta para determinações qualitativas e quantitativas de compostos moleculares diversos. Os espectros de absorção na região do infravermelho são relacionados às transições de uma molécula de um estado vibracional e/ou rotacional para outro. A radiação no infravermelho ocupa a região espectral com números de onda entre 12.800 $\boldsymbol{cm^{-1}}$ e 10 $\boldsymbol{cm^{-1}}$ no espectro eletromagnético. Essa região é usualmente dividida em infravermelho próximo (NIR, do inglês *Near Infrared*), médio (MIR, do inglês *Middle Infrared*) e distante (FIR, do inglês *Far Infrared*) (SKOOG; HOLLER; CROUCH, 2009). A Tabela 1 ilustra essa divisão da região espectral em função do número de ondas, comprimento de onda e da frequência.

Tabela 1: Regiões do infravermelho. Adaptado de SKOOG; HOLLER; CROUCH, 2009.

Região	Número de onda (cm^{-1})	Comprimento de onda (nm)	Frequência (Hz)
NIR	12.800-4.000	780-2.500	$3{,}8\ 10^{14}$ - $1{,}2\ 10^{14}$
MIR	4.000-200	2.500-50.000	$1{,}2\ 10^{14}$ - $6{,}0\ 10^{12}$
FIR	200-10	50.000-1.000.000	$6{,}0\ 10^{12}$ - $3{,}0\ 10^{11}$

Os sistemas moleculares absorvem radiação infravermelha em frequências (energias) discretas e depois a converte em energias vibracionais e rotacionais. No processo de absorção, a radiação infravermelha que se assemelha a frequência de vibração natural da molécula é absorvida, alterando a amplitude da vibração. Para que ocorra absorção da radiação infravermelha por uma molécula, deve haver uma mudança no seu momento dipolo decorrente das vibrações de estiramento e deformação da ligação, gerando então um campo elétrico capaz de interagir com campo elétrico da radiação (PAVIA et al., 2009; SILVERSTEIN et al., 2005). Dessa forma, moléculas diatômicas homonucleares, como $\boldsymbol{O_2}, \boldsymbol{N_2}, \boldsymbol{Cl_2}, \boldsymbol{H_2}$, etc., não absorvem no infravermelho, uma vez que nenhuma variação significativa do momento dipolo é obtida durante a vibração dessas moléculas (SKOOG; HOLLER; CROUCH, 2009).

As vibrações moleculares são divididas em estiramentos de ligação e deformações angulares. Uma vibração de estiramento é caracterizada pela variação na distância entre os átomos no eixo de ligação, e uma vibração de deformação é definida como a variação do ângulo entre duas ligações em um mesmo plano ou fora dele. Adicionalmente, as vibrações de estiramento e deformação podem ser simétricas ou assimétricas. A Figura 1 representa um esquema das diferentes vibrações que podem ocorrer em uma molécula triatômica angular.

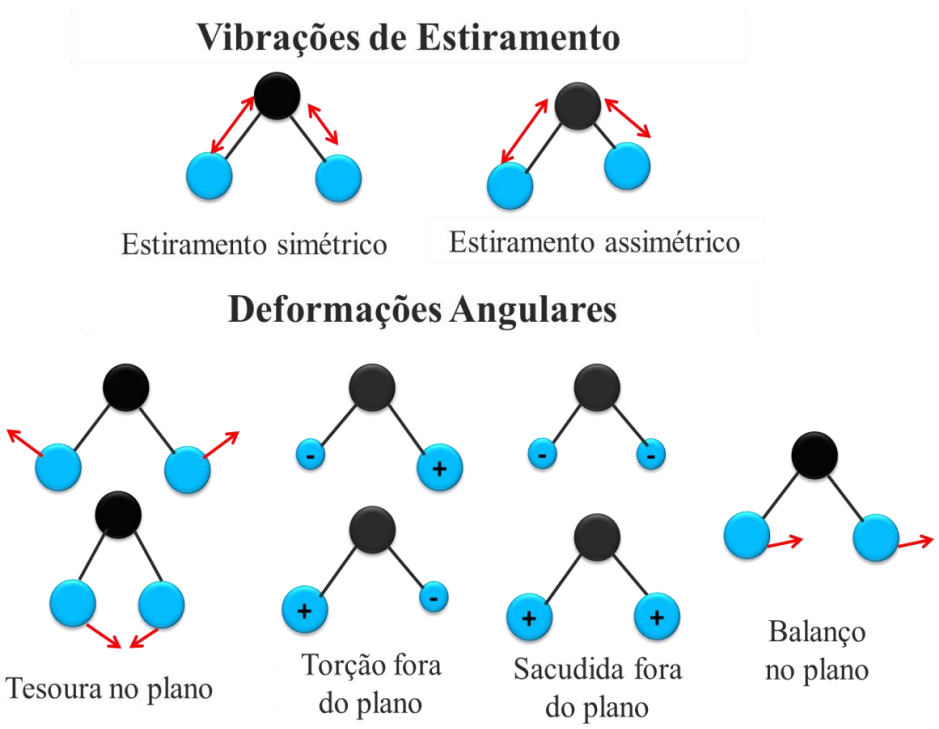


Figura 1: Estiramentos e deformações moleculares ativas no IR. Adaptado de SKOOG; HOLLER; CROUCH, 2009.

Uma maneira de interpretar as vibrações de estiramento nas moléculas é considerar que esse sistema se assemelha a um oscilador harmônico, em que duas massas (átomos) estão ligadas por uma mola (ligação covalente). Entretanto, esse modelo não é ideal para simular sistemas moleculares uma vez que fatores importantes que contribuem para as vibrações na molécula são negligenciados. O modelo do oscilador harmônico considera apenas vibrações entre dois níveis energéticos adjacentes ($\Delta\nu = \pm 1$) e a energia entre os níveis é sempre igual como pode ser visto na Figura 2A. Adicionalmente, dentro deste modelo as vibrações são independentes e não permitem haver combinações de vibrações e sobretons ($\Delta\nu = 2$ ou maior) (PASQUINI, 2003).

O modelo do oscilador anarmônico é mais adequado para descrever as vibrações em uma molécula. Diferente do oscilador harmônico, esse modelo considera os efeitos da repulsão coulombiana entre as nuvens eletrônicas dos átomos, quando eles se aproximam, no sentido restaurador da ligação química. Além disso, o modelo anarmônico prevê a redução da energia potencial devido ao decréscimo da força restauradora próximo da distância de dissociação da ligação. Para o oscilador anarmônico são permitidas transições fundamentais ($\Delta\nu = \pm 1$) e as transições com Δν=±2 ou 3 que explicam os sobretons, e também podem aparecer bandas de combinação de vibrações diferentes (SKOOG; HOLLER; CROUCH, 2009). A Figura 2B traz um diagrama simples que permite entender o funcionamento das transições em um modelo de oscilador anarmônico em função da energia potencial.

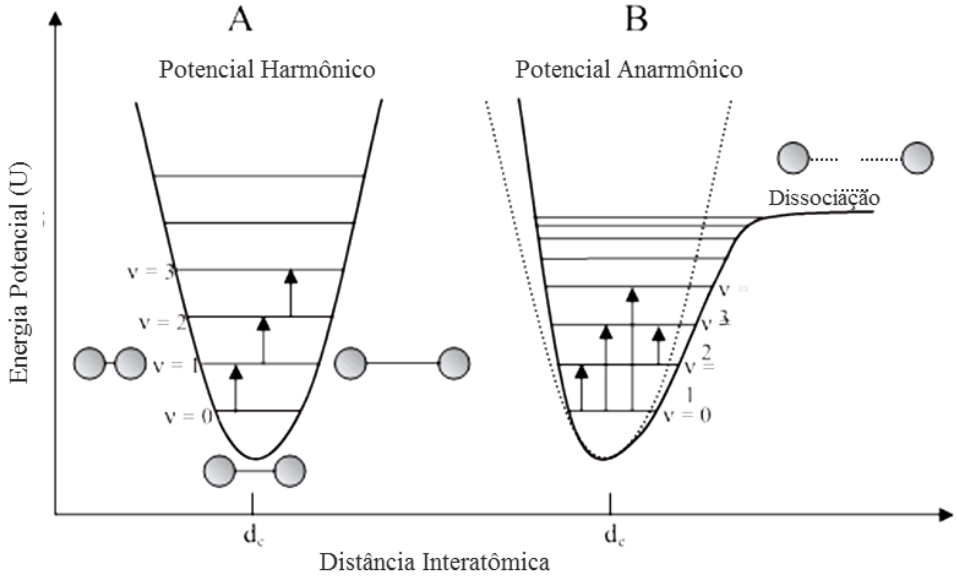


Figura 2: Diagrama de energia potencial para osciladores (A) harmônicos e (B) anarmônicos. Adaptado de (PASQUINI, 2003).

3.2 Espectroscopia no Infravermelho Próximo (NIR)

A espectroscopia NIR compreende a faixa entre os números de onda 12800 $\boldsymbol{cm^{-1}}$ e 4000 $\boldsymbol{cm^{-1}}$, na qual se observam os sobretons e bandas de combinação das vibrações fundamentais de $\boldsymbol{O-H}$, $\boldsymbol{N-H}$, $\boldsymbol{C-H}$ e $\boldsymbol{S-H}$. Os espectros são em geral obtidos por medidas de transmitância, refletância difusa ou transfletância. Atualmente a espectroscopia NIR é aplicada em rotina no controle de qualidade de fármacos, alimentos, polímeros, combustíveis e monitoramento de diversos processos industriais (PASQUINI, 2003).

O sistema de aquisição de HSI-NIR implementado na SisuCHEMA da Specim, que foi utilizada neste trabalho, se baseia na técnica de refletância difusa. Uma linha de fontes de radiação NIR realiza uma varredura da amostra e uma linha posterior de sensores acoplada à uma câmera coleta a radiação refletida. Uma grande área pode ser amostrada em um curto espaço de tempo, no entanto o número de detectores é fixo (MALINEN et al., 2010). Dessa forma, a largura da imagem tem um número de elementos (pixels) fixo. Para controlar a largura da área amostrada a câmera consta de um sistema de lentes capaz de selecionar linhas de largura específica (10 mm, 50 mm e 200 mm) e sempre com o mesmo número pixels.

3.3 Espectroscopia no Infravermelho Médio (MIR)

Uma das técnicas mais difundidas para elucidação e caracterização de compostos orgânicos é a espectroscopia de absorção no infravermelho médio. A espectroscopia MIR abrange a faixa espectral de 4000 $\boldsymbol{cm^{-1}}$ a 200 $\boldsymbol{cm^{-1}}$, que corresponde principalmente às vibrações fundamentais das moléculas. Alguns grupos químicos também apresentam picos de sobretons no espectro MIR (SILVERSTEIN et al., 2005; SKOOG; HOLLER; CROUCH, 2009). Em geral, para sólidos a espectroscopia MIR por Reflectância Total Atenuada (ATR, do inglês *Attenuated Total Reflectance*) mostra-se mais eficiente para obtenção de espectros com rapidez, de forma não destrutiva e com resolução e intensidade de sinal adequados. O acessório de ATR consiste em um cristal com alto índice de refração que fica em contato com a amostra e que permite a reflexão do feixe incidente entre seus planos. Os espectros são adquiridos quando um feixe de radiação infravermelha incide com um grau de inclinação específico no cristal e é refletido dentro do mesmo. Durante as múltiplas reflexões no cristal, a radiação MIR penetra na amostra com pequena profundidade, mas o suficiente para adquirir informações sobre a amostra (Figura 3). Essa radiação que é absorvida denomina-se *onda evanescente* (SKOOG; HOLLER; CROUCH, 2009).

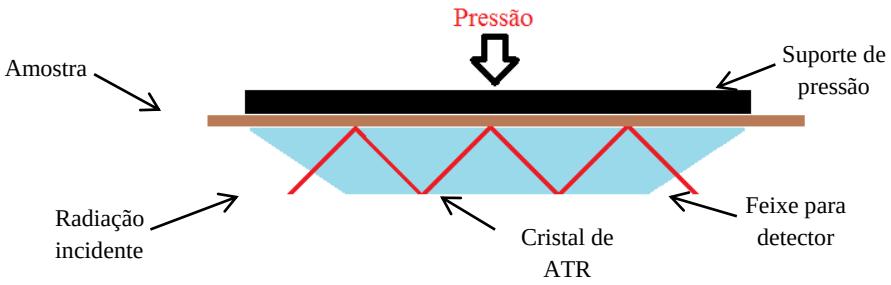


Figura 3: Esquema de aquisição espectral com acessório de ATR. Do autor.

A instrumentação necessária para aquisição de imagens com ATR é bastante recente. Em 2007, a Perkin Elmer introduziu o acessório para o instrumento Spotlight 300®, mas apenas em 2012 uma larga linha de acessórios de ATR-image foi disponibilizada (LING; SOMMER, 2015). A utilização de ATR na aquisição de imagens hiperespectrais no infravermelho médio (HSI-MIR) permite obter imagens com ótima resolução espectral e espacial. Um inconveniente da técnica é que o tamanho da imagem adquirida depende das dimensões do cristal utilizado e que ainda são restritas.

3.4 Imagens Hiperespectrais

Uma imagem digital pode ser definida como uma função em duas ou três dimensões, f(x,y), onde x e y e são coordenadas espaciais, e a amplitude de f para qualquer coordenadas xy é a intensidade no canal ou escala utilizado. Cada um desses conjuntos de coordenadas possuem localização e intensidade próprias e são denominadas de *pixel*. Assim, além das dimensões espaciais, as imagens digitais possuem uma dimensão adicional que pode assumir diferentes valores ou intensidades para cada pixel (GONZALES; WOODS, 2002; PRATS-MONTALBÁN et al., 2011).

Quando as imagens estão em escala de cinza cada pixel assume um único valor e, portanto, a matriz de dados que representa a imagem é bidimensional (Figura 4a). Por sua vez, as imagens RGB (do inglês *Red, Bue, Green*) são compostas por três canais de cores (vermelho, verde e azul) e sua matriz de dados é tridimensional, desse modo cada pixel assume três valores de intensidade, um para cada cor (Figura 4b)(PRATS-MONTALBÁN et al., 2011).

Existe um tipo especial de imagem que é capaz de fornecer informação química e espacial acerca da composição das amostras. Quando um pixel dessas imagens representa um espectro completo elas são chamadas de Imagens Hiperespectrais, ou imagens

multiespectrais, quando o pixel representa apenas alguns comprimentos de onda (Figura 4c). A matriz de dados de uma imagem multi/Hiperespectral é tridimensional e é geralmente denominada de cubo ou hipercubo (DE JUAN et al., 2009).

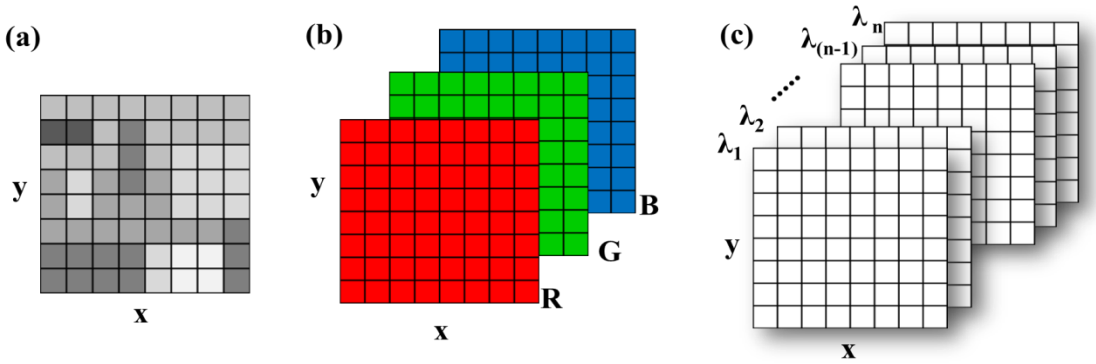


Figura 4: Matriz de dados das imagens (a)em escala de cinza, (b) em RGB e (c) Hiperespectrais. Do autor.

Em se tratando de imagens hiperespectrais, a coordenada espectral pode ser obtida por diferentes técnicas como espectroscopias nas regiões do Infravermelho, UV-Vis e Raman. Todas essas técnicas fornecem um grande volume de dados e importantes informações sobre a composição química e distribuição nas amostras. No entanto, uma análise direta das imagens pode não revelar as informações mais importantes e limitar o potencial da técnica de HSI. Nesse sentido, o uso de métodos quimiométricos é imprescindível para avaliar as imagens e processar os dados (DE JUAN et al., 2009).

Para realizar o pré-processamento de imagens hiperespectrais é necessário em geral desdobrar a matriz de dados em duas dimensões para que as ferramentas quimiométricas clássicas possam ser utilizadas. O hipercubo é desdobrado em uma matriz bidimensional, na qual os pixels ocupam as linhas e as variáveis espectrais ocupam as colunas da matriz. A decomposição do cubo ocorre de acordo com a Figura 5:

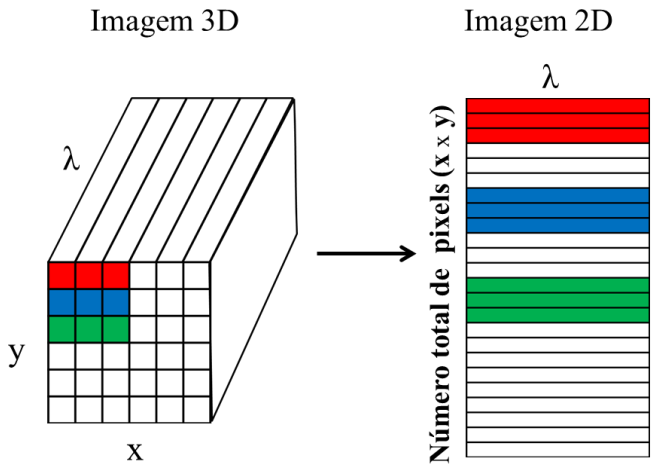


Figura 5: Desdobramento da imagem de 3D para 2D. Do autor

Com a matriz desdobrada em 2D é possível aplicar as técnicas de pré-processamento de espectros e depois remontar o hipercubo, o que permite recuperar as informações de distribuição espacial dos pixels.

3.5 Técnicas de Pré-processamento de Dados

Nos últimos anos o número de trabalhos que abordam aplicações HSI e quimiometria vem expandindo em diferentes áreas. Um dos catalisadores desse processo é o desenvolvimento de algoritmos e computadores que possibilitam tratar matematicamente as matrizes de dados das imagens hiperespectrais (VIDAL; AMIGO, 2012).

Dependendo do tipo de amostra e do equipamento utilizado para aquisição, as imagens podem guardar informações pouco relevantes ou até mesmo irrelevantes para o objetivo da análise. Um desses problemas de amostragem é a seleção de uma área maior que a região de interesse (ROI, do inglês *Regions of interest*), uma vez que as imagens adquiridas são sempre áreas retangulares e nem sempre a amostra tem esse formato. Outro problema na aquisição da ROI é o uso de acessórios que impedem a visualização direta da área amostrada, com isso pode haver a tomada de pixels que não correspondem a ROI. Um exemplo deste último caso é aquisição de imagens por meio de *ATR-image*. Com esse acessório a área da imagem adquirida depende do posicionamento da amostra sob o cristal e está sujeita a pequenos desvios.

Uma prática comum para seleção da ROI é usar a matriz dos escores de uma análise de componentes principais, que permite segmentar a imagem e, por consequência, facilitar a

seleção (VIDAL; AMIGO, 2012). É importante ressaltar que as diferenças entre a ROI e o restante da imagem nem sempre é clara e pode ser uma tarefa minuciosa. Vidal e Amigo (2012) sugerem três métodos para se realizar a seleção da ROI: por seleção manual; usando histogramas de frequência dos scores e por seleção de limites de valores nos escores da PCA. A seleção manual permite delimitar com muita precisão a ROI, por outro lado demanda muito tempo e é subjetiva, o que afeta a reprodutibilidade e dificulta a aplicação em larga escala. Já a seleção usando os histogramas e a determinação manual dos limites permite obter uma rápida e adequada seleção da ROI na maioria dos casos. No entanto, em algumas situações há um intervalo ambíguo de valores que podem pertencer tanto a ROI quanto ao restante da imagem. Nesta situação deve-se ponderar o que é mais vantajoso, garantir que apenas a ROI seja selecionada mesmo que custe a perca de parte da mesma ou selecionar toda a ROI e arriscar incluir alguma parte que não é relevante para o objetivo pretendido (VIDAL; AMIGO, 2012).

Quando se trabalha com HSI frequentemente aparecem pixels com características muito distintas dos demais pixels da imagem. Esses pixels anômalos ou *dead pixels*, podem derivar de flutuações atípicas nos detectores durante a aquisição das imagens que geram pontos, grupos e até linhas de pixels com valores nulos ou ausentes. Pixels com mais de 25% das variáveis nulas ou ausentes são considerados *dead pixel*. Esse critério, portanto, pode ser usado na busca por esses pixels numa imagem. Embora, alguns algoritmos de análise multivariada são capazes de lidar com um número limitado de *dead pixels*, esses pixels anômalos podem distorcer os modelos multivariados e até impossibilitar a criação dos mesmos (VIDAL; AMIGO, 2012). Dependendo da localização dos *dead pixels* na imagem, medidas preventivas podem ser tomadas para não prejudicar os modelos tais como a retirada de uma faixa da imagem que contém esses pixels, quando se localizam em zonas pouco relevantes da imagem, ou a correção dos pixels ruins, quando o corte não for uma opção viável. Uma forma de corrigir os *dead pixels* é substituí-los pela média ou mediana dos valores dos pixels vizinhos. A substituição dos pixels pela mediana é mais aconselhável que pela média, uma vez que evita a atribuição de valores muito altos (VIDAL; AMIGO, 2012).

A matriz de espectros resultantes do desdobramento de uma imagem pode ser submetida às técnicas quimiométricas usuais de pré-processamentos para corrigir efeitos indesejáveis e maximizar as informações relevantes. A presença de ruído, linha de base, espalhamento espectral, entre outros fenômenos indesejados na matriz de dados depende das características da amostra, da técnica de aquisição de imagem e dos acessórios utilizados. De modo geral,

espectros de infravermelho são bastante susceptíveis aos efeitos de espalhamento de radiação principalmente na espectroscopia por reflectância (RINNAN et al., 2009).

A normalização, a variação normal padrão (SNV, do inglês *Standard Normal Variate*), a correção do espalhamento multiplicativo (MSC, do inglês *multiplicative scatter correction*), a suavização com filtro de Savitzky-Golay (SG) e as derivadas espectrais são técnicas normalmente empregadas.

A normalização é aplicada para corrigir os efeitos da diferença de concentração ou quantidade de amostra. É uma técnica que opera nas amostras, onde cada variável de um vetor (ou espectro) é dividido por uma constante. Essa constante pode ser, por exemplo, a soma de todos os elementos do espectro (normalização pela área unitária) ou a raiz quadrada da soma quadrática dos elementos (normalização pelo comprimento unitário). Outra possibilidade é usar o valor absoluto máximo do espectro (normalização pelo máximo) como constante. Para espectros derivados é ideal usar a normalização pela faixa que divide cada variável espectral pela diferença entre o máximo e mínimo das variáveis (BEEBE et al., 1998; RINNAN et al., 2009).

A MSC é uma técnica capaz de corrigir efeitos aditivos e multiplicativos dos espectros resultantes do espalhamento da radiação. Cada espectro da matriz é corrigido com base em um espectro de referência que geralmente é o espectro médio. O primeiro passo da MSC é estimar os coeficientes de correção (b_0 e $b_{ref,1}$) através da Equação1 e por fim realizar a correção por meio da Equação2.

$$\boldsymbol{x_{org} = b_0 + b_{ref,1} \cdot x_{ref} + e} \quad \textbf{(1)}$$

$$\boldsymbol{x_{corr} = \frac{x_{org} - b_0}{b_{ref,1}} = x_{ref} + \frac{e}{b_{ref,1}}} \quad \textbf{(2)}$$

Onde: $\boldsymbol{x_{org}}$ é o espectro original, $\boldsymbol{x_{ref}}$ é o espectro de referência, $\boldsymbol{x_{corr}}$ é espectro corrigido, $\boldsymbol{e}$ é o resíduo deixado pela modelagem e $\boldsymbol{b_0}$ e $\boldsymbol{b_{ref,1}}$ são os coeficientes de correção associados aos efeitos aditivos e multiplicativos, respectivamente (RINNAN et al., 2009). Pelo fato da MSC utilizar um espectro de referência para corrigir os efeitos de espalhamento, o pré-processamento deve ser refeito no caso da remoção de alguma amostra anômala (em inglês: *outlier),* uma vez que esse procedimento implica numa mudança em $\boldsymbol{x_{ref}}$ e, por consequência, de $\boldsymbol{x_{corr}}$. SNV é uma técnica de normalização capaz de corrigir o espalhamento

da radiação tão bem quanto o MSC. Embora tenham o mesmo objetivo, as duas técnicas partem de princípios diferentes para realizar a correção de espalhamento. SNV atua individualmente em cada amostra sem a necessidade de um espectro/vetor de referência. O método de SNV é bastante simples: o espectro é centrado na média e em seguida é feita uma normalização usando o desvio padrão (RINNAN et al., 2009). A Equação 3 demonstra como os espectros são corrigidos.

$$\boldsymbol{x_{corr} = \frac{x_{org} - a_0}{a_1}} \qquad \textbf{(3)}$$

Onde: $\boldsymbol{x_{org}}$ é o espectro original, $\boldsymbol{x_{corr}}$ é espectro corrigido, $\boldsymbol{a_0}$ é o valor médio do espectro a ser corrigido e $\boldsymbol{a_1}$ é o desvio padrão do espectro. Uma das vantagens de se usar SNV é que ela não é influenciada pela remoção de *outlier* ou qualquer outro espectro.

Embora o pré-processamento com MSC e SNV tenham em geral resultados muito semelhantes, algumas vezes o efeito sobre a dispersão das amostras no gráfico de escores da PCA é muito diferente. O uso de SNV para pré-processamento de espectros de imagem no NIR pode resultar em um gráfico de scores com formato de elipse ou concha (Figura 6a). Quando a MSC é aplicada outro padrão de dispersão é adotado no gráfico de escores e as amostras tendem a evidenciar possíveis *outliers* (Figura 6b), (FEARN et al., 2009).

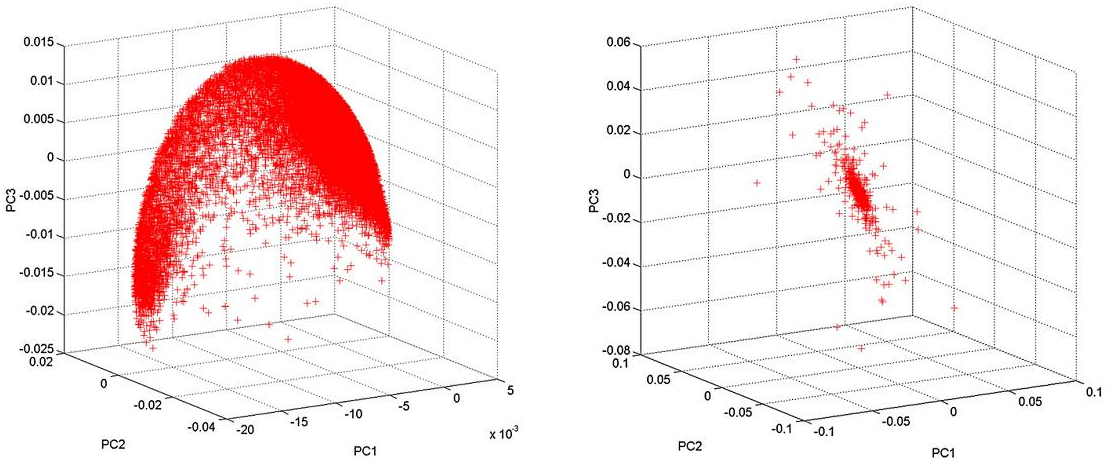


Figura 6: Gráfico de escores de PCA com (a) SNV e (b) MSC. Adaptado de FEARN et al., 2009.

Ambos os pré-processamentos, SNV e MSC, podem ser aplicados aos dados espectrais sem prejuízos aos modelos de PCA. No entanto, quando os efeitos da modelagem realizada com SNV são muito intensos no gráfico de escores, eles podem influenciar os modelos de calibração ou classificação, mais especificamente os modelos que assumem uma distribuição normal dos scores. A MSC por sua vez, pode influenciar a separação de um número maior de *outliers* que outras técnicas quando seus efeitos estão fortemente refletidos nos escores (FEARN et al., 2009).

A derivada espectral é outra técnica capaz de corrigir os efeitos de espalhamento de radiação que atua ao longo das variáveis (BEEBE et al., 1998). A derivada com filtro Savitzky-Golay é uma das técnicas de pré-processamento de espectros mais citadas em artigos científicos (LUO et al., 2005). Essa técnica tem uma etapa de suavização do espectro que consiste em ajustar um polinômio a uma janela de variáveis espectrais. Em seguida, uma derivada de qualquer ordem é aplicada (RINNAN et al., 2009). A derivada de Savitzky-Golay além de evidenciar os sinais pouco intensos, também pode maximizar ruídos dos espectros (BEEBE et al, 1998), por esta razão a definição do tamanho da janela de suavização (número de pontos usados para ajustar o polinômio) é uma etapa determinante para o sucesso no pré-processamento. A janela tem que ser grande o suficiente para corrigir o ruído espectral, mas não tão grande ao ponto de reprimir informações espectrais importantes.

No que se refere ao tratamento de imagens hiperespectrais, nem sempre um único tipo de técnica é capaz de fornecer os resultados esperados. Na maioria dos casos, as imagens passam tanto pelo pré-processamento espectral como espacial através das técnicas de processamento de imagens (VIDAL; AMIGO, 2012). Além dos pré-processamentos apresentadas aqui, diferentes técnicas estão disponíveis na literatura para aplicar em vários tipos de matrizes de dados com os mais variados objetivos.

3.6 Análise de Componentes Principais (PCA)

A Análise de Componentes Principais é uma técnica exploratória largamente aplicada que permite, entre outras coisas, reduzir a dimensionalidade dos dados e agrupar as amostras de acordo com a similaridade. A PCA cria um novo sistema de eixos ortogonais na direção de maior variância dos dados. Esses eixos são denominados de PC’s (do inglês *principal components*), (BEEBE et al., 1998). A Figura 7 representa um conjunto de amostras e duas variáveis no novo sistema de eixos das PC’s:

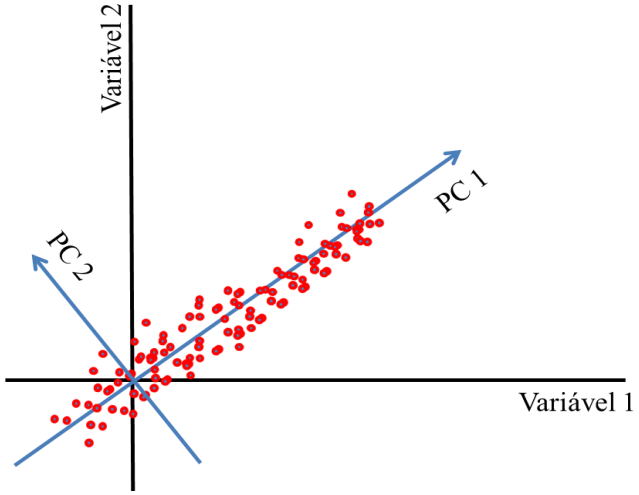


Figura 7: Gráfico da dispersão de amostras no sistema de eixos originais (eixos pretos) e nas duas primeiras PC's (eixos azuis). Adaptado de (BEEBE et al., 1998).

A PCA é realizada de modo que a primeira PC descreve o máximo de variância dos dados, a segunda PC descreve o máximo de variância restante e é ortogonal a primeira, e assim por diante para as próximas PC's. Técnicas espectroscópicas em geral, armazenam as informações sobre as amostras em muitas variáveis e quando associadas a técnicas de imagem os conjuntos de dados podem ultrapassar a casa dos milhões em número de variáveis como, por exemplo, as imagens hiperespectrais na região do infravermelho. Portanto, as técnicas como a PCA que é capaz de comprimir as informações relevantes num pequeno número de dimensões são muito importantes para avaliação dos dados.

Na PCA a matriz de dados após desdobramento da imagem (**X**) é decomposta em três matrizes: matriz de escores (**T**), que determina as coordenadas das amostras/pixels no novo sistema de eixos; matriz de *loadings* ou pesos (**P**), que representa a contribuição ou o peso de cada variável original para as novas variáveis; matriz de erros (**E**), que representa o resíduo deixado pelo modelo (WOLD; ESBENSEN; GELADI 1987). A Equação 4 representa matematicamente a decomposição da matriz de dados no novo sistema de eixos.

$$\mathbf{X} = \mathbf{T}.\mathbf{P}^{\mathbf{t}} + \mathbf{E} \qquad \textbf{(4)}$$

A Figura 8 é uma visualização esquemática da decomposição dos dados nas PC'se facilita o entendimento a cerca da compressão dos dados. A variável ***k*** *r*epresenta o número de PC's necessárias para explicar a variância dos dados.

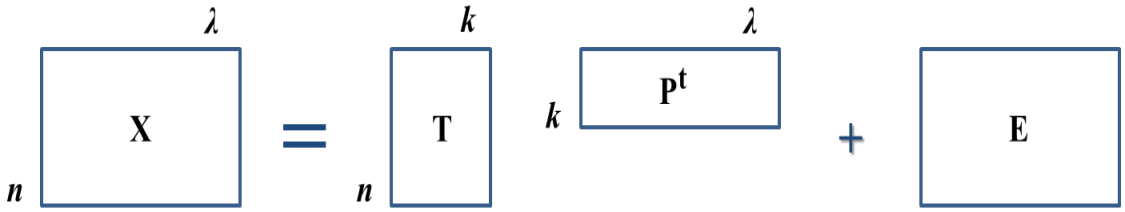


Figura 8: Esquema de decomposição da matriz **X** em escores e *loadings*. Adaptado de (BEEBE et al., 1998).

A matriz de escores **T** é a projeção das amostras no novo sistema de eixos, que resulta em um gráfico de dispersão das amostras (gráfico de escores) e permite avaliar o comportamento espacial dos dados em função da variância (BEEBE et al, 1998; WOLD; ESBENSEN; GELADI, 1987). A disposição espacial das amostras em função da variância pode ser de grande importância para determinar as semelhanças e/ou diferenças entre as mesmas.

No entanto, quando se trabalha com HSI, a dimensão $\boldsymbol{n}$ (número de pixels) da matriz **X** resulta do desdobramento da dimensão espacial $\boldsymbol{xy}$ do hipercubo da imagem (ver Figura 5). Como essa dimensão é preservada após a PCA, os pixels podem ser reagrupados para formar uma imagem de escores em função de cada uma das **PC** e, consequentemente, da variância dos dados.

3.7 Projection Pursuit (PP)

O desenvolvimento de técnicas multivariadas possibilitou obter muitas informações a cerca das características das amostras. Um conjunto de dados espectroscópicos, por exemplo, pode guardar inúmeras características, tais como o as absorções em determinados comprimentos de onda, a natureza do ruído, a relação entre amostras, a relação entre as variáveis, a largura de bandas, amostras anômalas, etc. (WENTTZEL; HOU, 2012). Todas essas características podem estar contidas no conjunto de dados, mas não estarem acessíveis através da abordagem de análise dos dados utilizadas. Outro fator importante é que muitas vezes o volume de dados obtidos é muito grande que torna difícil extrair a informação desejada de tal conjunto. Dessa forma, as ferramentas de análise multivariada dos dados deve possibilitar acessar as informações mais relevantes em busca das características desejadas, incluindo uma etapa de redução da dimensionalidade e avaliação das informações remanescentes.

Como já mencionado, a PCA é uma técnica exploratória que permite a visualização de dados de alta dimensão em um espaço de pequena dimensão sem perdas significativas de

informação. As informações sobre o conjunto de dados são extraídas em função da máxima variância para cada componente principal e reorganizadas no novo espaço de dados. Mas algumas vezes a variância pode não ser o melhor parâmetro para obter as informações desejadas (HOU; WENTTZEL, 2011). Nestes casos outros parâmetros podem oferecer projeções mais adequadas para avaliar os dados. A situação descrita pode ser mais bem compreendida observando a Figura 9, que representa amostras pertencentes a duas classes.

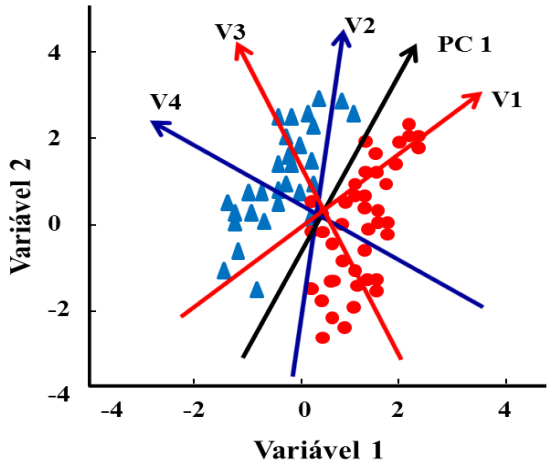


Figura 9: Comparação de uma projeção no sentido da máxima variância (PCA) e diferentes vetores de projeções interessantes, para amostras pertencentes a duas classes. Adaptado de (HOU; WENTZELL, 2011).

A Figura 9 mostra um conjunto de dados com diferentes vetores de projeção, entre eles está a PC1. Fica evidente na figura que a direção de projeção com máxima variância explicada não contém as informações úteis para discriminar as duas classes. Entre os outros vetores de projeção, o **V4** é responsável pela melhor discriminação. Portanto, faz-se necessário o uso de outras funções que sejam capazes de determinar a direção de projeção mais *"interessante"* para avaliar as características dos dados e assim acessar as informações mais relevantes para discriminar as classes (HOU; WENTZELL, 2013).

Em 1974 Friedman e Tukey apresentaram, pela primeira vez, PP como técnica exploratória. PP é um algoritmo de mapeamento linear que utiliza a distância entre pontos, assim como a variância, para buscar a forma mais interessante de projeção dos dados (FRIEDMAN; TUKEY, 1974). Assim, a finalidade da técnica é análoga a da PCA, buscando uma forma de representar os dados em um novo espaço de projeções com dimensões reduzidas e que preserve as características desejadas. Embora tenha o mesmo objetivo da PCA, *Project Pursuit* procura ajustar uma função, diferente da variância, que seja mais bem correlacionada com a propriedade de interesse. Essa função é denominada de *índice de*

projeção (HOU; WENTZELL, 2011). O algoritmo de PP associa cada uma das direções de projeção no espaço multidimensional a um *índice de projeção* capaz de medir o quão "interessante" é esta projeção para a descrição das características dos dados (FRIEDMAN; TUKEY, 1974).

Diferentes *índices de projeção* podem ser utilizados para acessar as informações do conjunto de dados e cada um desses índices tem que ser testado e avaliado para determinar a sua utilidade. A variância é uma das funções que pode ser usada como índice de projeção e, por consequência, a PCA é um tipo de Project Pursuit. Qualquer função que seja relacionada diretamente com a normalidade dos dados pode ser adotada, sendo que a maior parte dos índices de projeção é utilizada para medira não normalidade da distribuição. Essa é uma propriedade interessante, uma vez que frequentemente está relacionada com elementos da estrutura dos dados, tais como agrupamentos (*clusters)* e outliers (HOU; WENTZELL, 2011). A curtose (Equação 5) é um índice de projeção que vem se destacando, pois essa função é capaz de evidenciar agrupamentos e, em alguns casos, *outliers*.

$$\boldsymbol{k} = \frac{\mathbf{1}/\boldsymbol{n}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^{\mathbf{4}}}{(\mathbf{1}/\boldsymbol{n}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^{\mathbf{2}})^{\mathbf{2}}} \qquad \mathbf{5}$$

Onde $\boldsymbol{k}$ é a curtose, *n* é o número de objetos (amostras), $\boldsymbol{x}_{\boldsymbol{i}}$ é o valor do parâmetro para cada amostra e $\overline{\boldsymbol{x}}$ é a média dos valores de todas as amostras. Para um conjunto de dados genérico a distribuição dos objetos, em geral, obedece a uma distribuição normal onde os valores estão dispersos em torno de uma média ($\overline{\boldsymbol{x}}$). A curtose pode descrever matematicamente a dispersão dos valores, conforme mostrado na Equação 5 (HOU; WENTZELL, 2011). A Figura 10 mostra a distribuição normal dos objetos em torno da média $\overline{\boldsymbol{x}}$ com diferentes curtoses.

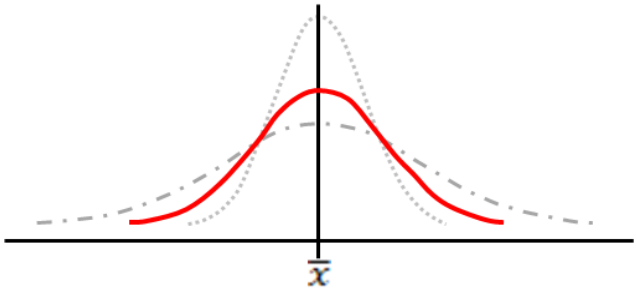


Figura 10: Representação esquemática da distribuição normal com diferentes curtoses. Adaptado de (WENTZELL et al., 2014).

Observando a distribuição dos objetos (ou amostras) em torno da média em uma direção apenas, pode-se identificar um ou mais agrupamentos de objetos, no entanto, esses agrupamentos podem estar mascarando outros conjuntos que se formam ao longo de outra direção de projeção. Outra direção de projeção é testada em busca de agrupamentos diferentes, e assim por diante (PEÑA; PRIETO, 2001a). Dessa forma, a distribuição normal é fragmentada sucessivas vezes à medida que projeções "interessantes" são propostas. Cada novo agrupamento encontrado se distribui em torno de uma média que pode ser descrita pela curtose. A partir da maximização ou minimização da curtose dos dados podem ser encontrados *outliers* ou agrupamentos, respectivamente. Conjuntos de dados que possuem uma distribuição bimodal tendem a ter pequenos valores de curtose, adicionalmente os valores se distribuem em torno de diferentes médias em cada máximo da curva normal. Dessa forma, minimizar a curtose implica em evidenciar o caráter bimodal da distribuição e separar agrupamentos em torno de novas médias. Por outro lado, maximizar a curtose implica em detectar amostras que se distanciam de uma distribuição normal dos dados como os *outliers* (PEÑA; PIETRO, 2001a). A Figura 11a mostra vetores cujas projeções possuem diferentes valores de curtose (V1, V2, V3 e V4) e variância (inclusive a PC1), para um conjunto de dados genérico. A Figura 11b apresenta uma representação das curtoses, evidenciando que o vetor V4 discrimina melhor as classes e, consequentemente, o comportamento bimodal.

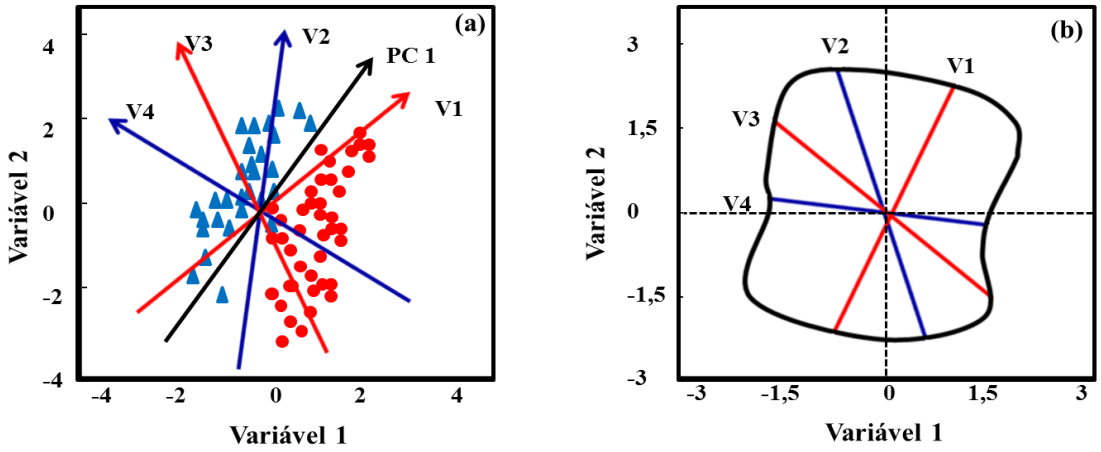


Figura 11: Gráfico de duas dimensões de dados genéricos: (a) representação de vetores cujas projeções mostram diferentes valores de curtose e variância; (b) representação dos vetores de curtose, onde a magnitude da curtose é determinada pela distância dos pontos à origem. Adaptado de (HOU;WENTZELL, 2011).

De fato, o índice de projeção curtose permite vários vetores de projeção para um conjunto de dados, porém aqueles vetores que possuem os menores valores de curtose são escolhidos para discriminação das classes (HOU; WENTZELL, 2011). Após otimizar o vetor com mínima curtose, qualquer variação que esteja correlacionada com ele é retirada do conjunto de dados e um segundo vetor de projeção é otimizado para a matriz de resíduos. Uma nova matriz de resíduos é obtida e novamente um vetor de projeção é otimizado, e assim sucessivamente até obter um conjunto de vetores suficiente para descrever as características de interesse. O procedimento descrito refere-se à determinação de *curtose univariada* e resulta em uma série de vetores de projeção ortogonais, capaz de preservar as principais características do conjunto de dados. É importante ressaltar que a variância dos dados é preservada ao longo de cada vetor projetado e também a sua posição relativa (HOU; WENTZELL, 2013). Dessa forma, PP também gera uma matriz de *loadings* e escores como a PCA.

Outra similaridade entre as duas técnicas é a situação limite para o modelo criado. Na PCA quando o número de PC's utilizada para explicar a variância dos dados é muito grande ocorre um sobreajuste do modelo. Para PP usando curtose univariada, a situação limite é atingida quando os objetos (ou amostras) são projetados nos vértices de um retângulo ou paralelepípedo, dependendo do número de dimensões do espaço de projeção.

Na curtose univariada os vetores de projeção são obtidos um por vez e em sequência, já com o algoritmo de *curtose multivariada* todos os vetores são otimizados simultaneamente. A diferença reside no fato de que a curtose multivariada busca o subespaço, um plano ou

hiperplano com duas ou três dimensões, de projeção como uma única entidade e a otimiza critérios diferentes (HOU; WENTZELL, 2011), ou seja, todos os vetores de projeção tomados como uma única função que será otimizada. A curtose multivariada também possui um caso limite, onde os objetos (ou amostras) estão igualmente distribuídos numa circunferência (ou elipse) para projeção bidimensional ou na superfície de um cilindro quando projetado em três dimensões (HOU; WENTZELL, 2013). A Figura 12a e b mostra o comportamento de um conjunto de dados genérico, projetado em duas dimensões, quando submetido à PP com minimização de curtose univariada e multivariada, na situação de caso limite, respectivamente.

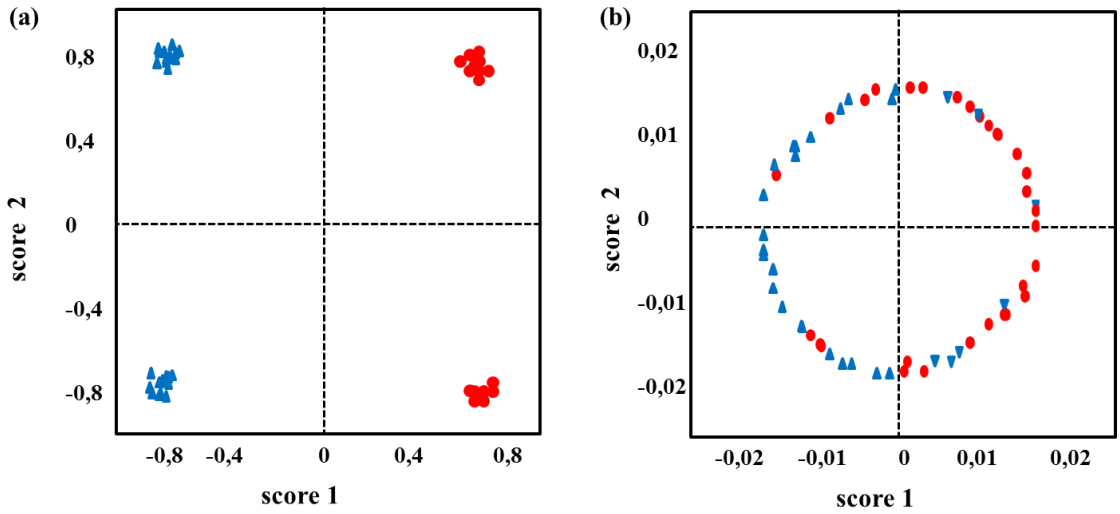


Figura 12: Gráfico bidimensional de dados genéricos com otimização de: (a) curtose univariada; (b) curtose multivariada. Adaptado de (HOU; WENTZELL, 2013).

Como já mencionado, o caso limite univariado resulta na projeção dos objetos nos vértices de um retângulo, ficando os constituintes de uma mesma classe separados devido ao uso de informações sem correlação estrutural (Figura 12a). Algo similar está representado pela Figura 12b, onde o caso limite de curtose multivariada levou à disposição dos objetos em uma circunferência. A situação limite resulta numa minimização extrema da curtose, onde a curva da distribuição normal é achatada ao ponto de fornecer estruturas que não estão correlacionadas com as características dos dados, mas são apenas correlações aleatórias (HOU; WENTZELL, 2013).

Em geral, a curtose univariada apresenta os melhores resultados para casos que se tem poucas classes com tamanhos próximos. Em casos com muitas classes e de tamanhos muito distintos o ajuste dos vetores de projeção pode se tornar problemático usando o algoritmo de

curtose univariada. Para esses casos é indicado o uso da curtose multivariada (HOU; WENTZELL, 2013).

A *Project Pursuit* é uma técnica com ótimo desempenho e capaz de obter uma maior quantidade de informações úteis que a PCA para os casos em que o número de amostras é significativamente maior que o número de variáveis, numa razão de no mínimo 8:1. Entretanto, na química analítica, especialmente na espectroscopia, é comum que o número de variáveis supere o número de amostras. Os casos limite são um reflexo do número insuficiente de amostras para otimizar os vetores de projeção e reduzir a curtose (HOU; WENTZELL, 2013). Em outras palavras, o número elevado de variáveis força o agrupamento das amostras e seu deslocamento para o limite de dispersão no gráfico de escores (para curtose uni e multivariada). Assim, embora PP possa ser usada como ferramenta de compactação da dimensionalidade, o método utilizado para tal não lida bem com números elevados de variáveis e necessita de uma técnica para redução de variáveis.

Tendo em vista essa limitação de PP, Hou e Wentzell (2011) propuseram a PCA como método para redução do número de variáveis. Em vez de usar a totalidade das variáveis, podem-se usar as variáveis latentes ou PC's. No entanto, se um número muito grande de variáveis latentes for utilizado pode haver problemas de sobreajuste e se for muito pequeno pode haver perdas significativas das informações. Como o desempenho de PP depende do nível de compressão (número de PC's) e da quantidade de informação preservada, faz-se necessária uma metodologia capaz de determinar o número mais adequado de PC's.

Recentemente, Wentzell e colaboradores (2015) propuseram uma maneira de avaliar o nível de compressão dos dados antes da implementação de PP que se baseia na a comparação dos espaços de projeção obtidos com número de PC's diferentes e sequenciais. Números de PC's adjacentes e que contém informações significativas devem apresentar projeções semelhantes. No entanto, comparar esses espaços de projeção não é uma tarefa trivial e demanda um método para fazê-la (WENTZELL et al., 2015).

3.7.1 Análise de Procusto (PA)

Tendo em vista as dificuldade para comparar dois espaços de projeção, Wentzell et al. (2015) demonstram que a Análise de Procusto (PA, do inglês *Procrustes Analysis*) pode ser usada para avaliar o grau de similaridade entre dois espaços obtidos com diferentes níveis de compressão.

Essa técnica busca o ajuste linear dos objetos em um novo espaço com mesmas dimensões com o intuito de avaliar a similaridade da disposição espacial dos mesmos em duas representações com diferentes números de PC's. O segundo espaço criado é submetido a diferentes operações matemáticas (como, por exemplo, rotação e reflexão), com o objetivo de aproximar ao máximo os dois espaços e estimar matematicamente a similaridade entre os espaços (WENTZELL et al., 2015). Essa estimativa é feita através de um coeficiente de dissimilaridade que pode ser encarado como o efeito dos resíduos deixados da aproximação entre um dos espaços e sua estimativa criada a partir do outro espaço de projeção.

Uma representação de como ocorre à análise de Procusto pode ser vista na Figura 13, onde dois conjuntos de dados quaisquer $\boldsymbol{X}$ e $\boldsymbol{Y}$ com dimensões $\boldsymbol{nxp}$ foram submetidos a uma PCA para obtenção das matrizes de escores $\boldsymbol{Z}$ e $\boldsymbol{T}$. Em seguida as matrizes de escores foram submetidas a PA de modo que $\boldsymbol{Z}$ deva ser ajustada a $\boldsymbol{T}$. Após uma série de operações matemáticas (rotação, translação, reflexão e centralização) o melhor ajuste entre $\boldsymbol{T}$ e $\boldsymbol{Z}$ foi obtido.

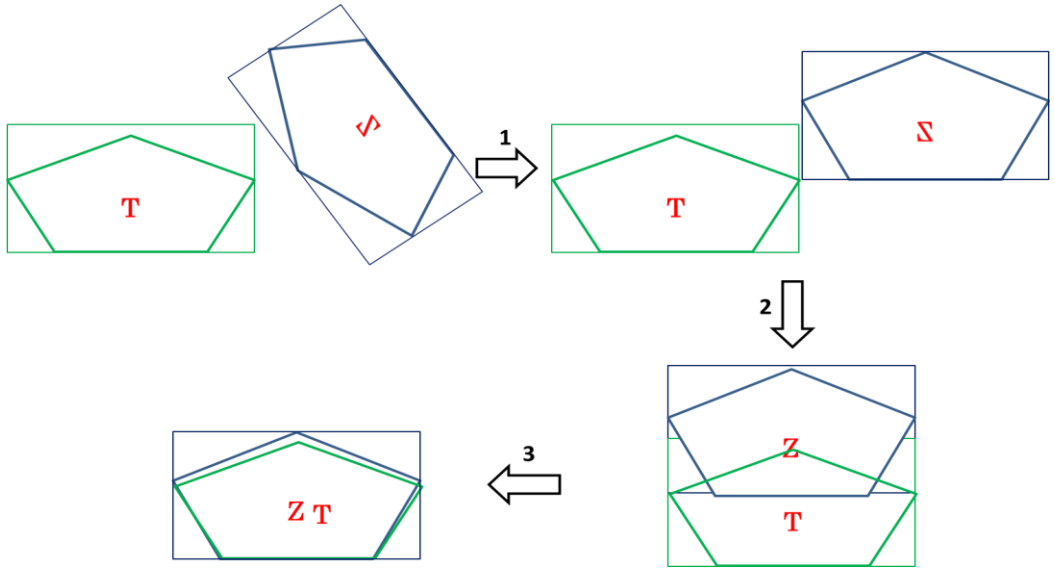


Figura 13: Esquema de ajuste dos espaços em uma análise de procusto através das operações de rotação (1), translação seguida de reflexão (2) e centralização (3). Do autor.

O esquema da Figura 13 mostra que as duas matrizes de escores apresentam uma pequena falta de ajuste, que pode ser calculado através do *coeficiente de dissimilaridade* (d) usando a Equação 6. Em termos matemáticos d é calculado como a soma quadrática das distâncias de cada elemento de $\boldsymbol{Z}$ e seu correspondente em $\boldsymbol{T}$ normalizada pela variância de $\boldsymbol{Z}$ em torno de sua média (Equação 7). Dessa forma, quanto menor a distância entre cada um dos elementos

correspondentes de $\boldsymbol{Z}$ e $\boldsymbol{T}$, menor é o coeficiente d e maior é a similaridade dos espaços (WENTZELL et al., 2015).

$$\boldsymbol{d} = \frac{\sum_i \sum_j \left(\boldsymbol{z_{ij}} - \hat{\boldsymbol{z}}_{\boldsymbol{ij}}\right)^2}{\sum_i \sum_j \left(\boldsymbol{z_{ij}} - \bar{\boldsymbol{z}}_{\boldsymbol{j}}\right)^2} \qquad \mathbf{6}$$

$$\boldsymbol{Z} = \boldsymbol{aTR} + \boldsymbol{B} + \boldsymbol{E} = \hat{\boldsymbol{Z}} + \boldsymbol{E} \qquad \mathbf{7}$$

Onde $\bar{\boldsymbol{z}}_{\boldsymbol{j}}$ é a soma dos valores $\boldsymbol{j}$ em $\boldsymbol{Z}$ e as variáveis $\boldsymbol{z_{ij}}$ e $\hat{\boldsymbol{z}}_{\boldsymbol{ij}}$ são elementos das matrizes de Z e $\hat{Z}$, respectivamente. R e B são as operações de rotação e translação, respectivamente, e E é a matriz dos resíduos.

Para realizar PP em dados comprimidos com PCA é necessário utilizar análise de procusto para garantir que os vetores de projeção obtidos carreguem as informações que lhes permitam representar o máximo de características significativas do conjunto de dados original. Essas projeções não devem ser sensíveis a pequenas variações no processo de compressão e diferentes projeções estáveis podem ser obtidas à medida que o número de PC's varia (WENTZELL et al., 2015). Quando o máximo de informações significativas é extraído pelos diferentes vetores de projeção restam apenas as informações sem correlações significativas, cuja combinação linear em suas variáveis originam diferentes vetores de projeção instáveis.

No entanto, muitas vezes os conjuntos de dados são adquiridos através de uma medida que fornece mais de um tipo de informação sobre as amostras, como, por exemplo, espectros de infravermelho contendo diversas bandas de absorção muito intensas ao longo de todo o espectro. Para esse tipo de matriz de dados podem surgir mais diferentes projeções estáveis que representam as informações extraídas (WENTZELL et al., 2015). Tais características dos dados podem ser interpretadas mais facilmente através do *mapa de procusto*.

O *mapa de procusto* permite avaliar as diferenças entre os vetores de projeção estáveis e instáveis, identificar regiões com projeções similares e determinar as condições que produzem as diferentes projeções estáveis e os diferentes padrões de agrupamento. A Figura14 apresenta um mapa de procusto e alguns gráficos de escores obtidos através da análise de procusto de dados submetidos a PP, usando PCA para comprimir os dados.

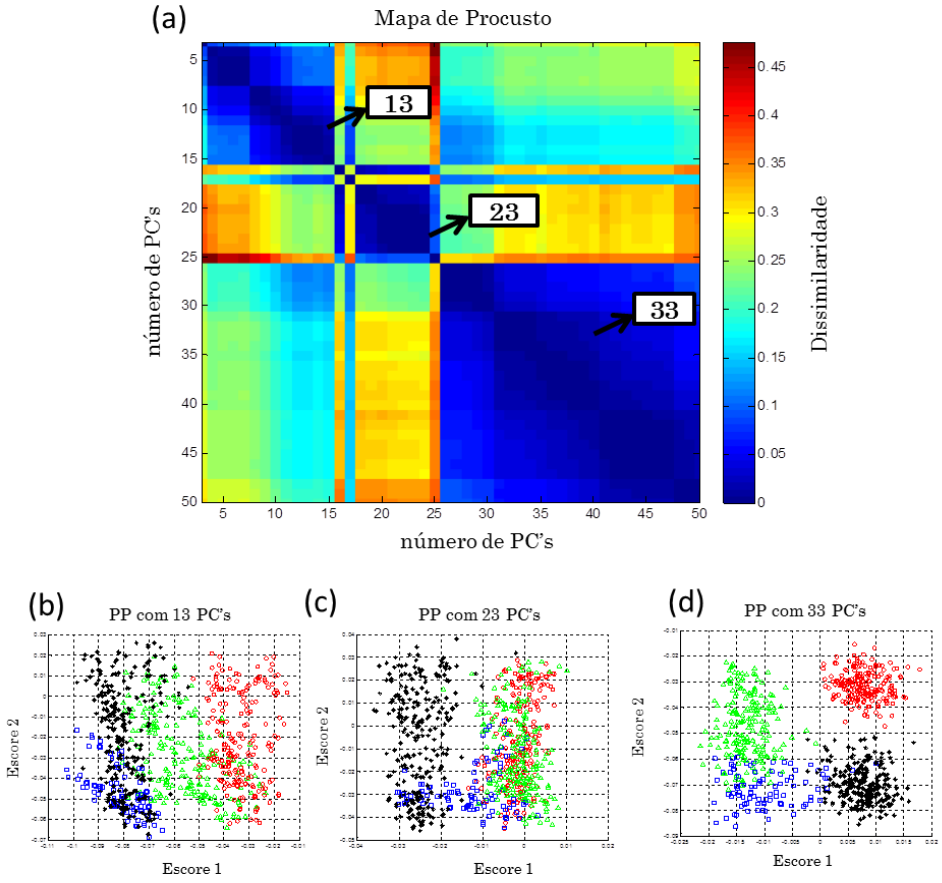


Figura 14: (a) Mapa de procusto de PP, com compressão de dimensionalidade, para quatro grupos de amostras genéricas e os respectivos gráficos de escores de dois vetores extraídos da PA com auxílio do mapa usando (b) 13 PC's, (c) 23 PC's e (d) 33 PC's. Do autor.

O mapa de Procusto permite avaliar as semelhanças entre os vetores de projeção obtidos em diferentes níveis de compressão por PCA, variando de 3 a 50 PC's. Do mapa de Procusto, diferentes vetores de projeção podem ser extraídos, de modo que em cada ponto do mapa dois ou três vetores em um nível de compressão específico são selecionados para compor o novo espaço de projeção. No exemplo acima foram extraído vetores em três níveis de compressão, o primeiro nível com 13 PC's, o segundo com 23 e o terceiro com 33 (Figura 14b, c e d). Em cada nível de compressão os vetores permitiram avaliar a discriminação dos agrupamentos. Outra característica muito importante do mapa da Figura 14a é que ele possui três grandes regiões de estabilidade, caracterizadas pelo pequeno valor de dissimilaridade apresentado em escala de cores (WENTZELL et al., 2015).

Tendo em vista toda a discussão sobre a análise de PP, percebe-se que essa técnica tem grande potencial como ferramenta para análise exploratória de dados espectroscópicos. Além

disso, pelo fato de gerar uma matriz de escores, PP pode ser aplicada às imagens hiperespectrais para obter imagens dos escores.

3.8 Imagens Hiperespectrais e Investigação Forense

A utilização de metodologias científicas para ajudar na investigação policial é conhecida como investigação forense e as diferentes áreas do conhecimento utilizadas são denominadas como ciências forenses. A utilização das metodologias científicas é de grande importância para a investigação criminal, pois permite recompor e inferir sobre os fatos ou cenários investigados (BRAZ, 2014). Em seu cotidiano, os cientistas forenses se deparam com diferentes situações e a metodologia empregada para investigá-las deve ser objetiva, confiável e específica para cada situação, visto as implicações de suas conclusões.

Dessa forma, as ciências forenses podem ser fragmentadas em diversos campos de estudo para melhor atender as necessidades da investigação policial. A *Documentoscopia* é uma dessas linhas de investigação. Ela é responsável por avaliar documentos a fim de atestar sobre sua autenticidade. Muitas pesquisas foram desenvolvidas e outras estão em desenvolvimento para auxiliar nesta e nas demais áreas da ciência forense. As técnicas de imagens hiperespectrais têm ganhado espaço nas investigações forense e já possui diversas aplicações, abordando diferentes situações da investigação criminal.

Chen e colaboradores (2009) desenvolveram um trabalho para determinação de impressões digitais e um contaminante em laminas de alumínio, por meio de imagens hiperespectrais na região do infravermelho médio (HSI-MIR). Os autores preparam uma superfície limpa de alumínio revestido e depositaram uma impressão digital. Na sequência, outra impressão contaminada com uma solução de um explosivo (RDX) foi depositada sobre a primeira simulando uma situação de sobreposição de digitais. Os autores utilizaram a decomposição em valores singulares para identificar qual das digitais continha o RDX associando os picos de maior absorção do explosivo com a sua distribuição espacial na imagem dos escores. Eles demonstraram que é possível identificar o histórico químico de uma impressão digital, por meio de HSI-MIR, quando a mesma encontra-se sobreposta a outra impressão digital.

Li e colaboradores (2013) utilizaram imagens hiperespectrais na região do visível e análise discriminante para estimar a idade de manchas de sangue de até 30 dias. Os autores produziram 10 amostras em uma folha de papel branco, cada amostra consistia de duas gotas

de sangue depositadas separadamente para criar duas manchas de sangue. Em seguida realizaram medidas de reflectância na região do espectro visível ($400 - 720\ nm$) diariamente durante 30 dias nas duas manchas de sangue de cada amostra, totalizando 620 imagens. Uma mancha foi usada para criar um conjunto de calibração e a outra para criar um conjunto de teste. Os espectros foram pré-processados com SNV. Para criar o modelo LDA os autores utilizaram apenas 5 comprimentos de onda de cada espectro. Encontraram uma razão de classificação correta (RCC) de 60,7% com erro médio de $\pm 1{,}17$ dias para os 30 dias. Eles demonstraram que a acurácia do modelo era maior em períodos de tempo menores. Para os primeiros 7 dias o RCC foi de 83,9% com erro médio de $\pm 0{,}27$ dias. Para os últimos 15 dias os autores obtiveram um RCC menor (53,5%) com erro médio maior ($\pm 1{,}62$).

No campo da Documentoscopia o uso de HSI ainda não tem muitas aplicações. Os trabalhos desenvolvidos nesta área, em sua maioria, utilizam técnicas subjetivas e algumas vezes destrutivas. A utilização de imagens na análise documentos na maioria dos casos se limita a imagens em escala de cores, fontes de iluminação e inspeção visual. Em se tratando diretamente da análise das tintas utilizadas para produzir os documentos, diversos autores deram sua contribuição para o desenvolvimento da Documentoscopia.

Um dos primeiros trabalhos que avaliou a problemática de cruzamento de traços foi desenvolvido por Igoe e Reynolds (1982). Para determinar a ordem da sobreposição de tintas em um cruzamento de tintas de caneta ball-point, os autores utilizaram papel fotográfico KromeKote para extrair parte da tinta do documento duvidoso. Para fazer a determinação de qual traço estava por cima, os autores consideraram que a concentração de tinta na borda da linha é maior que no centro, desse modo, a linha grafada primeiro perderia essa borda quando cruzada por outra linha. Igoe e Reynolds depositaram o papel fotográfico sobre o texto com o cruzamento de traços e variaram três fatores: pressão, temperatura e abrasão. Os autores obtiveram 80 % de determinação correta nos casos em que os traços tinham um ponto de intersecção e 100 % de acerto quando havia duas intersecções. Esse resultado não pode ser reproduzido para tintas líquidas, pois o princípio do método não se aplica a esse tipo de tinta.

Quando se trata de análise de documentos para fins forenses, o cuidado habitual com a integridade das amostras é redobrado, pois em muitos casos o documento tem valor financeiro e como prova. Durante uma investigação criminal é provável que os testes sejam refeitos para confirmar o resultado, portanto, técnicas como a de Igoe e Reynolds poderiam comprometer o valor judicial do documento. São necessários métodos objetivos e não destrutivos, que

possibilitem manter a integridade do documento. Nesse sentido, as técnicas espectroscópicas se apresentam como promissoras, pois permitem obter bastante informação química dos documentos com rapidez e de forma não destrutiva. Vários trabalhos foram publicados sugerindo metodologias para análise de tintas em documentos por meio de espectroscopia UV/vis, espectroscopia Raman e por infravermelho (SILVA et al., 2014b; SILVA et al., 2014c; THANASOULIAS; PARISIS; EVMIRIDIS, 2003; ADAM; SHERRATT; ZHOLOBENKO, 2008; SILVA et al., 2013). Alguns pesquisadores relatam o uso de imagens hiperespectrais em documentoscopia. É o caso dos trabalhos de Braz et al. (2015) que utilizou imagens Raman, Bojko et al.(2008) com imagens no infravermelho médio e Silva et al.(2014a) que usaram imagens no infravermelho próximo.

Silva e colaboradores (2014b) propuseram aplicar análise discriminante por mínimos quadrados parciais (PLS-DA, do inglês *partial least squares for discriminant analysis*) em espectros na região do visível obtidos por um equipamento comparador espectral de vídeo (VSC, do inglês *video spectral comparator*) para discriminar tintas de caneta preta. Os autores utilizaram 55 canetas de marcas e modelos diferentes para preparar traços manuscritos em papel branco (marca chamex) e em seguida foram adquiridos espectros na faixa de $400 - 1000\ nm$, com $1\ nm$ resolução (600 variáveis), 11,8 de aumento e diafragma em 60%. Os espectros foram pré-processados com Savitzky-Golay, polinômio de 2ª ordem, janela de 11 pontos e centralização na média. Cerca de 60% dos espectros foram utilizados para construir os modelos PLS-DA e 40% para validar. A primeira análise mostrou que todas as canetas foram discriminadas. Os modelos PLS-DA criados foram utilizados para resolver testes-cego e um caso real de documento questionável. Todos os testes-cego foram resolvidos corretamente e o caso real foi solucionado corretamente. Silva e colaboradores (2014c) aplicaram uma metodologia similar para discriminar canetas de tinta azul de diferentes marcas, tipos de tinta e modelos. Nesse segundo estudo os autores obtiveram resultados similares para discriminação das canetas e para os testes-cego, no entanto para o caso forense real não foi obtida resposta conclusiva. Visto que a variabilidade de canetas (marcas, modelos e tipo de tinta) é consideravelmente maior que a utilizada na pesquisa, o desafio maior para a abordagem apresentada é construir um banco de dados mais representativo para criar modelos PLS-DA robustos.

Com o objetivo de revelar o comportamento de tintas de caneta em cruzamentos de traços, Braz e colaboradores (2015) avaliaram o potencial da técnica de imagem Raman para avaliação do cruzamento de linhas produzido com canetas azuis. As intersecções de traços

foram realizadas com 6 canetas com tintas de composição diferentes, com diferentes tempos de separação entre a aplicação das duas tintas e tipos diferentes de papel. O microscópio de imagem Raman da ThermoScientific, modelo DXR™xi foi utilizado para adquirir imagens com área de $600\ \mu m\ x\ 600\ \mu m$, que recobre o cruzamento e parte das linhas individuais, na faixa espectral de $85\ -\ 3500\ cm^{-1}$. Os espectros foram pré-processados com correção de linha de base usando polinômio de 4ª ordem. Na sequência a técnica de Resolução de Curvas Multivariadas (MCR, do inglês *multivariate curve resolution*) com duas componentes foi aplicada nas imagens. Braz et al.(2015) puderam determinar em muitas amostras a sequência correta de sobreposição das tintas. O padrão de sobreposição só não foi obedecido por uma das canetas, cuja ordem de aplicação ficou invertida nos cruzamentos contendo essa caneta. O autor também notou a formação de pontos na imagem dos cruzamentos que ele atribuiu à mistura das tintas. Os autores também demonstraram que a diferença no tempo de deposição das tintas não influenciou na determinação da ordem da sobreposição.

Bojko e colaboradores (2008) utilizaram imagens hiperespectrais na região do infravermelho médio, adquiridas usando um acessório de ATR, para determinar a sequência de deposição de tintas em cruzamentos de traço. Os cruzamentos foram preparados através da combinação de canetas de tinta preta de diferentes marcas e modelos e impressoras a laser e a jato de tinta. Além de combinar traços de diferentes instrumentos gráficos os autores também adquiriram imagens dos cruzamentos sob diferentes condições como a pressão aplicada na escrita e o tempo de espera para aquisição de novas imagens durante 12 meses. Avaliando apenas alguns comprimentos de onda característicos, os autores foram capazes de identificar a sequência de traços apenas nos casos em que as canetas esferográficas foram combinadas com a tinta de impressora a laser usando toner.

Silva e colaboradores (2014a) avaliaram o potencial de imagens hiperespectrais no NIR para determinação de fraudes em documentos. Utilizando 10 canetas de tinta preta foram produzidas amostras simulando falsificações por obliteração de texto e cruzamento de linhas. Para os casos de adição de texto foram usadas 16 canetas. Foram produzidas 90 amostras de obliteração de texto sobrepondo uma tinta sobre um texto curto previamente preparado com uma tinta diferente. As falsificações por adição de texto foram formuladas combinando tintas de caneta que não permitiam discriminação visual direta, totalizando 22 amostras. As amostras de cruzamento de traço consistiam em produzir linhas com as 10 canetas e depois imprimir uma segunda linha cruzando a primeira, usando uma impressora a laser e dois toners diferentes. Também foram produzidos cruzamentos usando as canetas para grafar linhas

cruzando a linha dos toners, o que totalizou 40 amostras de cruzamento. Foram adquiridas imagens hiperespectrais na faixa de $928 - 2524\,nm$ para todas as amostras e depois de aplicar diferentes pré-processamentos espectrais as imagens de obliteração e adição de texto foram submetidas à PCA e MCR-ALS (MCR-ALS, do inglês *multivariate curve resolution – alternating least squares*). As amostras de cruzamento de traço foram submetidas à MCR-ALS e PLS-DA. Os autores obtiveram sucesso na identificação de 39 casos de obliteração de texto, dentre as 90 amostras, usando PCA e MCR-ALS. Entre os 22 casos de adição de texto, 18 adulterações foram corretamente identificadas (82% de acerto) e apenas 4 amostras não forneceram resultados conclusivos. Para os cruzamentos de traço os autores consideraram que a presença de falhas ou *gaps* no mapa de distribuição do toner indicavam que a tinta de caneta foi aplicada antes do toner e conseguiram determinar corretamente a sequência de cruzamento em 17 das 20 amostras (85% de acerto).

A utilização de métodos que não necessitam informações prévias sobre o objeto avaliado é uma das principais demandas numa investigação forense real. Portanto, existe uma necessidade de desenvolver métodos capazes de obedecer a este critério. Em quimiometria existem ferramentas de análise exploratória de dados, como a PCA e PP, que tem grande potencial para auxiliar em investigações deste tipo. Para simular condições reais de uma investigação forense basta criar um cenário experimental que leve em consideração amostras de natureza desconhecida pelo analista para realizar testes cegos. Os testes cegos consistem em aplicar a metodologia desenvolvida para analisar amostras com composição desconhecida para o analista.

4 MATERIAIS E MÉTODOS

Neste trabalho foram abordados dois problemas da Documentoscopia. O primeiro problema se refere aos casos de adição de texto, em que um registro é feito com uma caneta e na sequência outra caneta é usada para adicionar um registro. O segundo problema é a situação em que um texto feito com uma caneta é adulterado por adição de um registro feito com outra caneta sobre o texto original. A Figura 15 abaixo é uma representação das duas situações-problema descritas anteriormente.

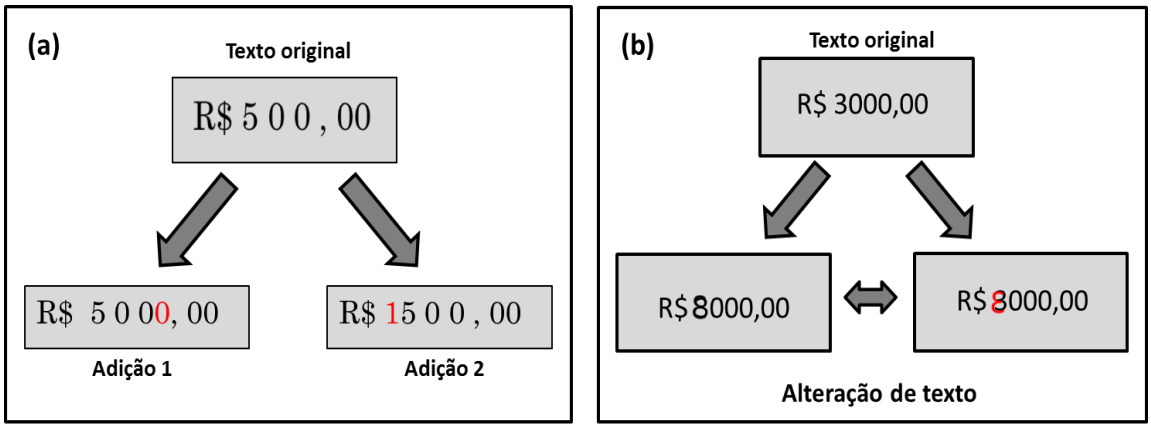


Figura 15: (a) Simulação de falsificação por adição de texto e (b) simulação de falsificação por adulteração do texto. Do autor.

A abordagem desses problemas foi realizada em duas etapas. Na primeira etapa buscou-se discriminar as diferentes canetas utilizadas para fazer os registros usando a técnica imagens hiperespectrais na região MIR. Na segunda etapa as técnicas HSI-NIR e HSI-MIR foram aplicadas em amostras preparadas para testes-cego com o objetivo de avaliar e validar os métodos utilizados na etapa discriminante e determinar seu potencial de aplicação para análises de fraudes em documentos reais.

4.1 Preparação de Amostras

Para preparar as amostras foram adquiridas 16 canetas de tinta preta entre as mais comuns disponíveis no mercado (SILVA et al., 2014a): 9 canetas Esferográficas (1 da marca Bic, 2 da Pentel, 1 Compactor, 1 Paper Mate, 1 Faber Castel, 1 Pilot, 1 Cello e 1 Mitsubishi), 3 canetas Gel (1 Bic, 1 Pentel e 1 Shneider), 2 canetas hidrográficas (1 Paper Mate e 1 Staedtler) e 2 canetas hidrográfica do tipo Rollerball (Bic e Pentel). As canetas foram codificadas com a primeira letra indicando o tipo da caneta, seguido de um número para ordenar as diferentes marcas, portanto os códigos das amostras são E1-9, G1-3, H1-2 e R1-2.

Os registros utilizados para análise discriminante foram preparados em pedaços de papel, com aproximadamente 3 cm² de área, provenientes de uma mesma folha de papel sulfite tamanho A4. Uma parte das amostras de teste-cego foi preparada em papel de cheque e a outra parte em pedaços de papel sulfite.

Utilizando uma caneta por vez, foram feitas linhas em dezesseis pedaços do papel sulfite, aplicando apenas uma camada de tinta e com pressão normal de escrita. Essas amostras foram utilizadas para realizar a análise de discriminação de tintas.

As amostras de teste-cego foram preparadas por três colaboradores (codificados como A, C e F) fazendo combinações duas a duas das dezesseis canetas fornecidas sem o conhecimento do avaliador. Entre as amostras produzidas, além de amostras simulando as fraudes já mencionadas, podia haver amostras sem qualquer adulteração. Dessa forma existia a possibilidade de identificar amostras com adição ou adulteração de texto e também amostras genuínas produzidas com apenas uma caneta.

Um colaborador preparou 10 amostras em uma mesma folha de cheque que consistiam de números com até 5 dígitos. Os outros dois colaboradores produziram mais 20 amostras em pedaços de papel sulfite, que consistiam em números de até dois dígitos. As amostras foram codificadas com as iniciais de Teste Cego e o código de cada colaborador mais o número da amostra, ou seja, TCA1-14, TCC1-6 e TCF1-10. Fotografias de todas as amostras estão apresentadas no Apêndice A

4.2 Aquisição de Imagens

Para a etapa de discriminação das canetas foram adquiridas imagens apenas na região do infravermelho médio. Para os testes-cego foram tomadas imagens no infravermelho próximo e médio.

4.2.1 Aquisição de HSI-MIR

Para cada amostra de traço preparada foi adquirida uma Imagem Hiperespectral na região do Infravermelho Médio (HSI-MIR) utilizando o Microscópio Spotlight 400 acoplado ao espectrômetro Spectrum 400 (Perkin-Elmer), usando um acessório de imagem por ATR. A faixa espectral foi de $4000 - 750\ cm^{-1}$, com dimensão fixa da imagem de $100\ x\ 100\ \mu m$, 128 scans, resolução espectral de 16 cm^{-1} e tamanho do pixel 6,25 µm. Dessa forma, cada

imagem obtida continha 256 pixels. Como o acessório de ATR tem uma pequena área de amostragem, apenas uma parte da linha foi amostrada.

As mesmas condições de aquisição descritas acima foram adotadas para as amostras do teste-cego, entretanto, o número de aquisições em cada amostra varia com as suspeitas sobre a amostra. Para cada amostra também foram adquiridas imagens do papel em que o número foi escrito para auxiliar no tratamento dos dados. A Figura 16 mostra alguns exemplos de como foram feitas as aquisições em cada amostra, de acordo com suas especificidades.

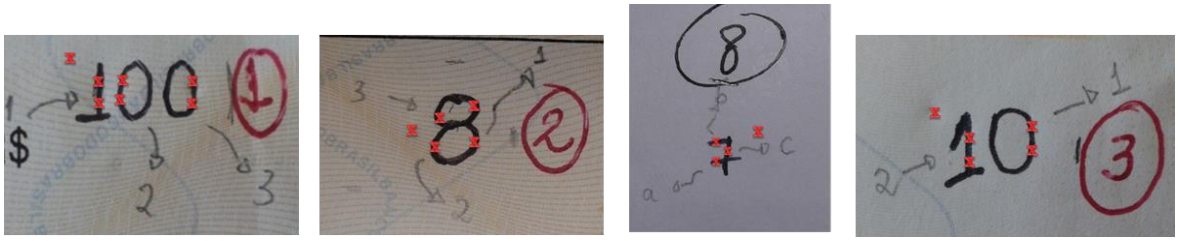


Figura 16: Método de aquisição de HSI-MIR. Os pontos em vermelho representam o local de aquisição de cada imagem. Do autor.

4.2.2 Aquisição de HSI-NIR

Para aquisição das Imagens Hiperespectrais no Infravermelho Próximo (HSI-NIR) foi utilizada uma câmera SisuCHEMA da Specim, com macrolente de captura de 10 mm. A faixa de trabalho foi de $928 - 2524\ nm$ com resolução espectral de $6{,}3\ nm$ e espacial de $10\ \mu m$. O tamanho da imagem variou entre $0{,}4$ e $1{,}5\ cm$ dependendo do tamanho do número a ser analisado. O tamanho do pixel foi mantido em $30\ x\ 30\ \mu m$. O número de pixels varia de acordo com o tamanho da imagem, mas é importante ressaltar que a macrolente adquire imagens com largura fixa de $10\ mm$ e com no máximo 320 pixels de largura.

4.3 Pré-Processamento dos Dados

As imagens HSI-MIR foram submetidas a diversos pré-processamentos para corrigir efeitos indesejáveis como o ruído e o espalhamento da radiação. Para o pré-processamento das imagens obtidas foram avaliadas as técnicas de SNV, MSC, normalização, suavização e 1ª derivada com filtro Savitzky-Golay. A suavização e a derivada foram realizadas usando polinômio de 2º grau e variando de 7 a 21 o número de pontos da janela. Nas imagens das amostras do teste-cego, além das técnicas de pré-processamento espectrais citadas, foi realizada a seleção da ROI por meio dos histogramas de frequência e imagem dos escores da PCA. A ROI foi adotada apenas para o teste cego para selecionar o máximo de pixels relacionados com a tinta.

O pré-processamento dos dados de HSI-NIR foi realizado empregando SNV, suavização com filtro SG e 1ª e 2ª derivada também com filtro SG. A suavização e as derivadas foram realizadas usando polinômio de 2º grau e variando de 7 a 13 o número de pontos da janela.

Para os dois tipos de conjunto de dados (HSI-MIR e HSI-NIR) o melhor pré-processamento foi determinado em função dos resultados da PCA e da correção no perfil dos espectros.

4.4 Análise dos Dados

As dezesseis imagens HSI-MIR obtidas das linhas feitas com cada caneta foram combinadas duas a duas a fim de obter o maior número de combinações distintas, totalizando 120 pares diferentes. Cada combinação da imagem de duas canetas foi submetida à PCA e PP.

Para cada amostra do teste-cego, após remover os pixels mais correlacionados com o papel de cada HSI-MIR (após a ROI), os conjuntos de dado das imagens da amostra foram submetidos à PCA e PP. Para o teste-cego usando as Imagens Hiperespectrais no NIR, PCA e PP foram aplicadas após o pré-processamento espectral e a imagem dos escores foi avaliada.

4.5 Softwares e Algoritmos

Inicialmente as imagens hiperespectrais no NIR foram cortadas e convertidas do formato *raw* para o formato *mat* usando o programa Evince 2.7.0 (UmBio). Depois de obtidos os cubos das imagens no NIR e as matrizes de dados no MIR os cálculos e pré-processamentos foram realizados usando o software MATLAB R2010a 7.10.0.499 (MathWorks).

A seleção de ROI foi realizada usando a rotina desenvolvida por Vidal e Amigo (2012) para seleção da ROI usando os histogramas de frequência da PCA, que faz parte. Os algoritmos desenvolvidos por eles também foram utilizados para realizar o pré-processamentos e a PCA nos conjuntos de dados de MIR e nas imagens hiperespectrais no NIR.

A análise de Project Pursuit e de Procusto foram realizadas através dos algoritmos desenvolvidos por Hou e Wentzell (2013) e Wentzell e colaboradores (2015), respectivamente.

5 ANÁLISE DISCRIMINANTE DE CANETAS DE TINTA PRETA

Foi determinado o melhor pré-processamento dos dados e em seguida as amostras foram submetidas à PCA e Project Pursuit. Embora esta última técnica não seja influenciada pelo tipo de pré-processamento espectral aplicado (HOU; WENTZEL, 2013), as amostras não podem ser submetidas a ela diretamente devido ao grande número de variáveis. Antes de usar PP, a dimensionalidade dos dados é reduzida usando PCA que é bastante sensível às técnicas de pré-processamento de dados e, portanto, os resultados aplicando PP também são influenciados.

5.1 Pré-Processamento Espectral na Região do MIR

Depois de obtidas as imagens hiperespectrais na região do MIR foi realizado o pré-processamento espectral. Como o objetivo principal da primeira etapa deste trabalho é a discriminação entre duas tintas diferentes, o melhor pré-processamento foi determinado usando os conjuntos de dados de duas canetas com tinta a base de óleo (canetas esferográficas), que apresentaram pior separação dos conjuntos de dados aplicando a PCA nos dados brutos. Para exemplificar as etapas que resultaram no melhor pré-processamento, os conjuntos de dados E1 e E3 serão adotados para demonstração, porém é importante ressaltar que foram observados os resultados de várias outras combinações de tinta antes de definir o melhor pré-processamento.

A Figura 17 apresenta os espectros brutos de cada pixel das imagens E1 e E3, excluindo apenas uma banda entre 2408 e 2240 cm^{-1} que representa apenas o CO_2 residual, após correção automática pelo software do equipamento. Observando os espectros brutos, pode-se notar a presença de efeito aditivo e bastante ruído nas duas primeiras bandas, que já era esperado devido ao uso de ATR. A região espectral entre 2900 cm^{-1} e 3500 cm^{-1} é caracterizada pelos estiramentos de $O-H$ e $C-H$, que são bandas características de celulose. Outras bandas importantes são encontradas em torno de 1580 cm^{-1} (estiramento $-C=C-$), entre 1300 cm^{-1} e 1500 cm^{-1} (deformação angular $C-H$) e entre 1160 cm^{-1} e 1030 cm^{-1} (estiramento de $C-O$).

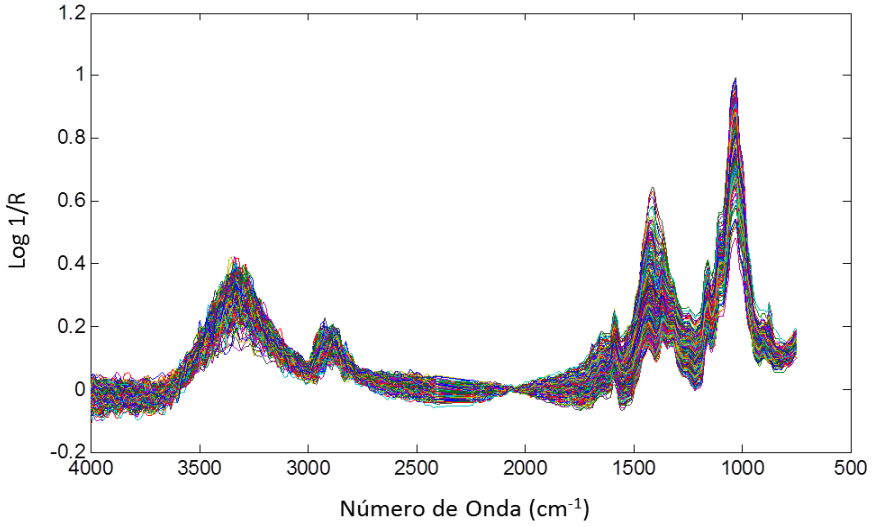


Figura 17: Espectros brutos das amostras E1 e E3.

O pré-processamento foi realizado em cada amostra separadamente com cada ferramenta de correção espectral. Inicialmente foi avaliado o efeito de SNV e MSC para os espectros e na discriminação de tintas e conclui-se que os resultados obtidos aplicando MSC são melhores na correção do efeito aditivo (Figura 18a-b) e também apresenta uma discriminação entre os conjuntos um pouco melhor (Figura 18c e d).

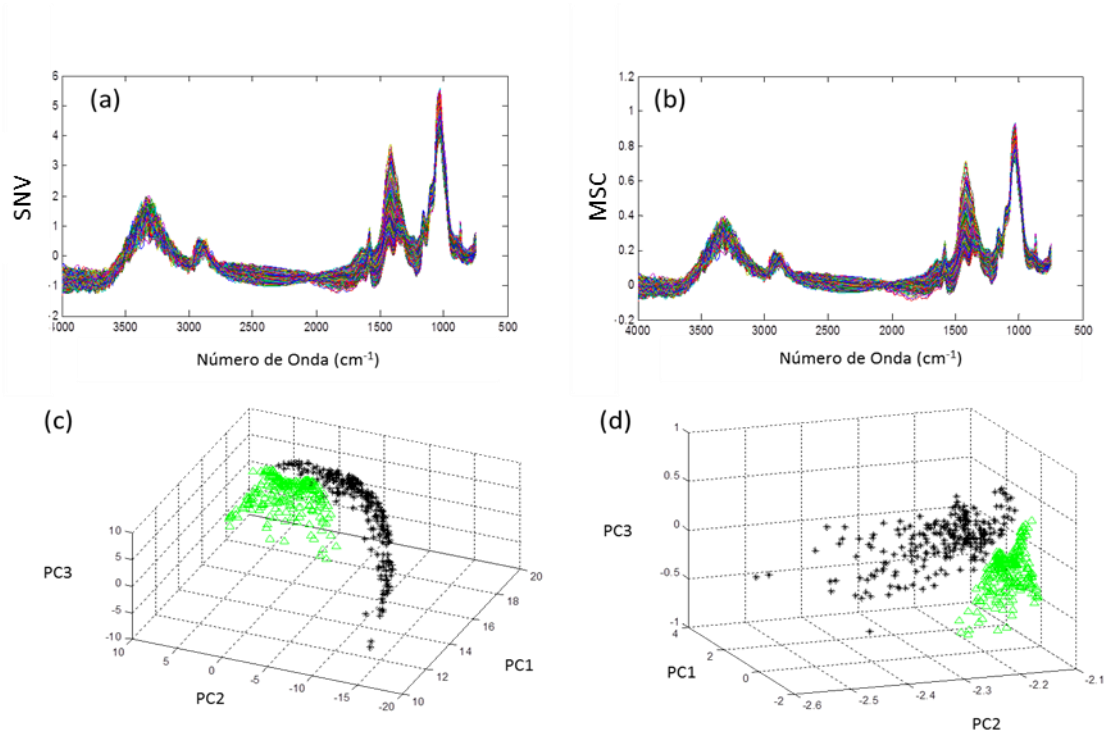


Figura 18: (a) Espectros das amostras E1 e E3 pré-processados com SNV e (b) com MSC. Gráfico dos escores da PCA para os espectros pré-processados (c) com SNV e (d) com MSC.

Nota-se na Figura 18c o efeito de projeção das amostras detalhado por Fearn e colaboradores (2008). Como será visto a seguir, os dados pré-processados com MSC também são afetados pelo efeito matemático na projeção dos dados. A suavização também foi testada associada à SNV e MSC, mas os resultados mantiveram-se inalterados em termos da discriminação dos grupos.

O outro pré-processamento avaliado foi a suavização seguida de derivada com filtro SG variando o tamanho da janela, que apresentou uma redução considerável do espalhamento espectral bem como do ruído. A Figura 19a-c mostra o efeito na redução do ruído e do espalhamento usando 1ª derivada com filtro de suavização SG e janelas de 7 a 11 pontos.

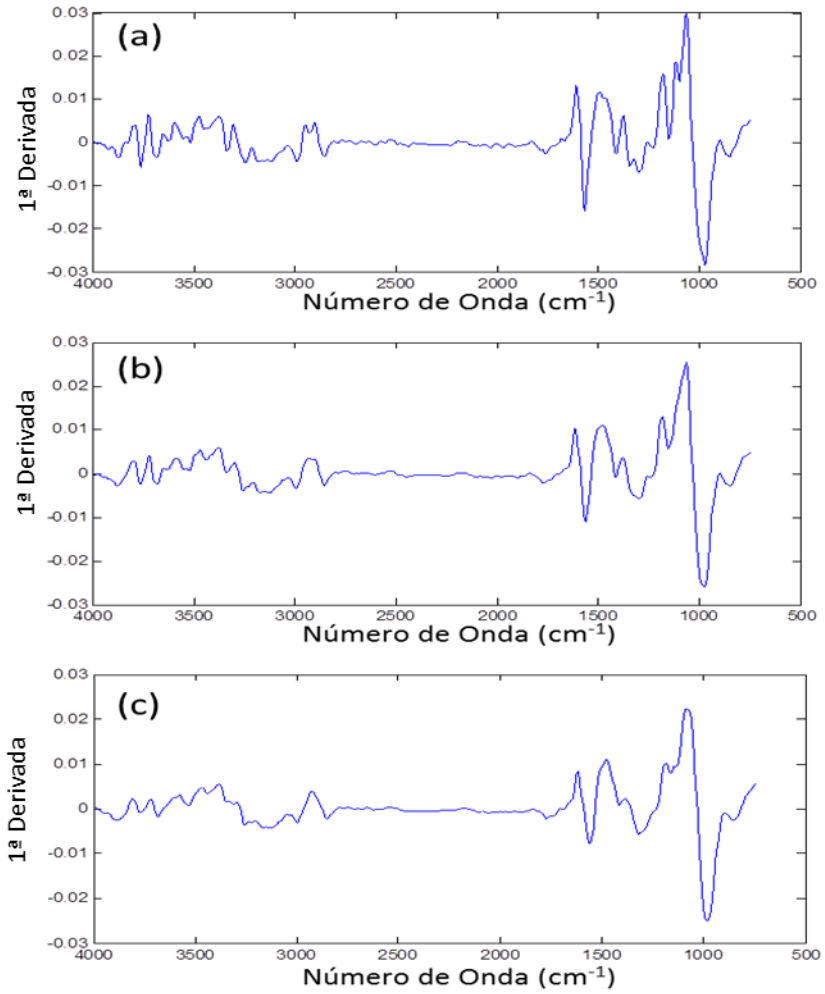


Figura 19: Espectro médio das amostras E1 e E3 pré-processadas com 1ª derivada e suavização com filtro SG, polinômio de segunda ordem e janelas de (a) 7 pontos, (b) 9 pontos e (c) 11 pontos

Nota-se que usando a suavização com janela de 7 pontos o espectro médio pré-processado ainda apresenta bastante ruído. Por outro lado, usando uma janela de 11 pontos

algumas bandas perdem resolução e o espectro torna-se menos informativo. Dessa forma, a janela de 9 pontos foi definida como a mais vantajosa para ser aplicada aos espectros. A 2ª derivada também foi testada, porém ela adicionou muito ruído aos espectros e não foi capaz de maximizar as informações mais relevantes. Depois de pré-processados, os conjuntos de dados foram submetidos à PCA e o resultado pode ser visto na Figura 20.

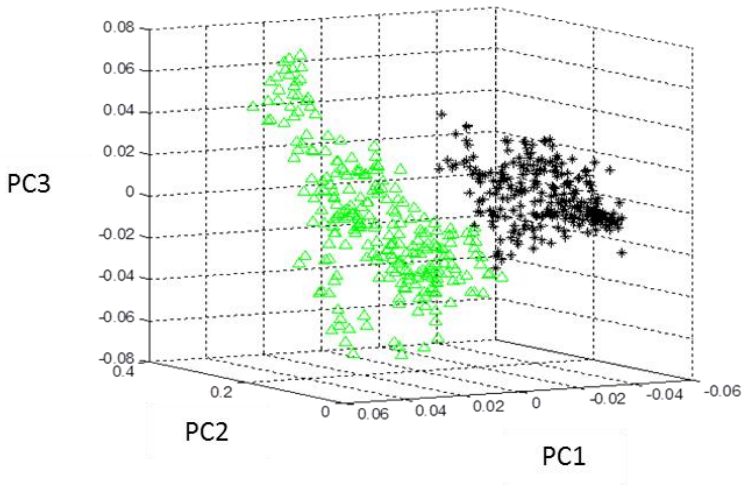


Figura 20: Gráfico dos escores da PCA das amostras E1 e E3 usando 1ª derivada com filtro SG, polinômio de segunda ordem e janelas de 9 pontos.

A princípio, a PCA dos dados pré-processados com MSC apresenta resultado um pouco superiores que a PCA nos dados pré-processados com derivada, porém o efeito do MSC na projeção das amostras no gráfico de escores pode dificultar a interpretação da PCA em alguns casos, uma vez que a dispersão em um eixo é maximizada apenas pelo efeito matemático da projeção. A Figura 21 é apenas outra visualização do gráfico de escores da PCA apresentado na Figura18d e mostra como os dados se alinham em torno de um eixo devido em consequência da correção feita pela MSC e, portanto, não reflete nenhuma característica química dos dados. Esse efeito foi detectado para todos os conjuntos de dados utilizados neste trabalho.

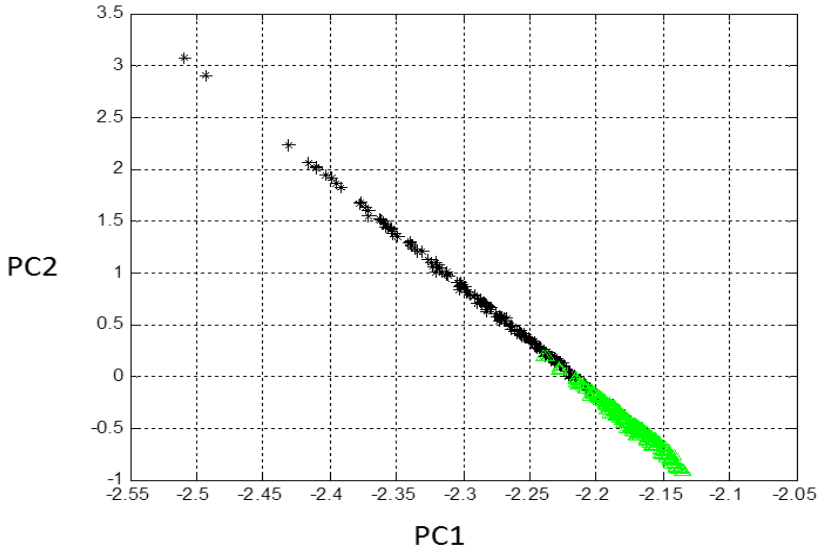


Figura 21: Gráfico dos escores da PCA das amostras E1 e E3 pré-processadas com MSC.

Tendo em vista que a diferença entre a discriminação obtida pela PCA usando derivada e MSC não é muito expressiva, optou-se pela 1ª derivada com filtro de suavização SG (polinômio de segunda ordem e janela de 9 pontos) para pré-processar todas as amostras.

Como as amostras preparadas podem conter quantidades diferentes de tinta em cada registro, a normalização dos dados foi realizada após a derivada. Assim, a normalização pela faixa foi testada para os conjuntos de espectros. Como pode ser visto na Figura 22, a normalização não forneceu nenhuma melhoria na discriminação das amostras.

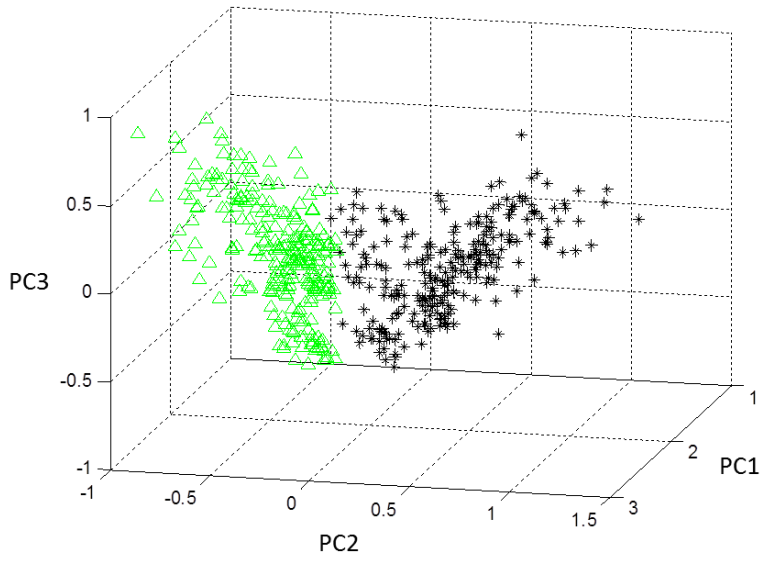


Figura 22: Gráfico dos escores da PCA das amostras E1 e E3 pré-processadas com 1ª derivada usando filtro de suavização SG (janela de 9 pontos e polinômio de 2ª ordem) e normalização pela faixa espectral.

Com base nos gráficos de escores das Figuras 20 e 22, podem-se formular duas hipóteses, a primeira é que a concentração da tinta não está influenciando na discriminação das mesmas e a segunda é que as concentrações de tinta usadas para produzir as amostras são muito semelhantes. A primeira hipótese é mais aceitável, uma vez que a quantidade de tinta em cada registro varia com o tipo e marca de caneta.

Depois de ponderar os efeitos de cada ferramenta de pré-processamento nos espectros e avaliar seu impacto para discriminação de tintas, foi estabelecida a 1ª derivada com filtro de suavização SG usando janela de 9 pontos e polinômio de segunda ordem como melhor pré-processamento para os conjuntos de dados das imagens hiperespectrais MIR.

5.2 Análise Discriminante das Canetas Usando PCA e Project Pursuit

As tintas de canetas esferográficas a base de óleo foram as que apresentaram maior dificuldade para serem discriminadas entre si e também de outros tipos de tinta. Entre as 36 combinações de canetas a base de óleo (canetas esferográficas), 15 não puderam ser discriminadas apenas com PCA. Quando a técnica de Project Pursuit foi aplicada apenas a combinação de E2 com E6 não formou dois grupos distintos. O APÊNDICE B apresenta uma tabela que reúne os resultados da análise discriminante para combinações de tintas com mesma base química. A Figura 23 representa alguns exemplos dos resultados obtidos usando a PCA para discriminar as tintas.

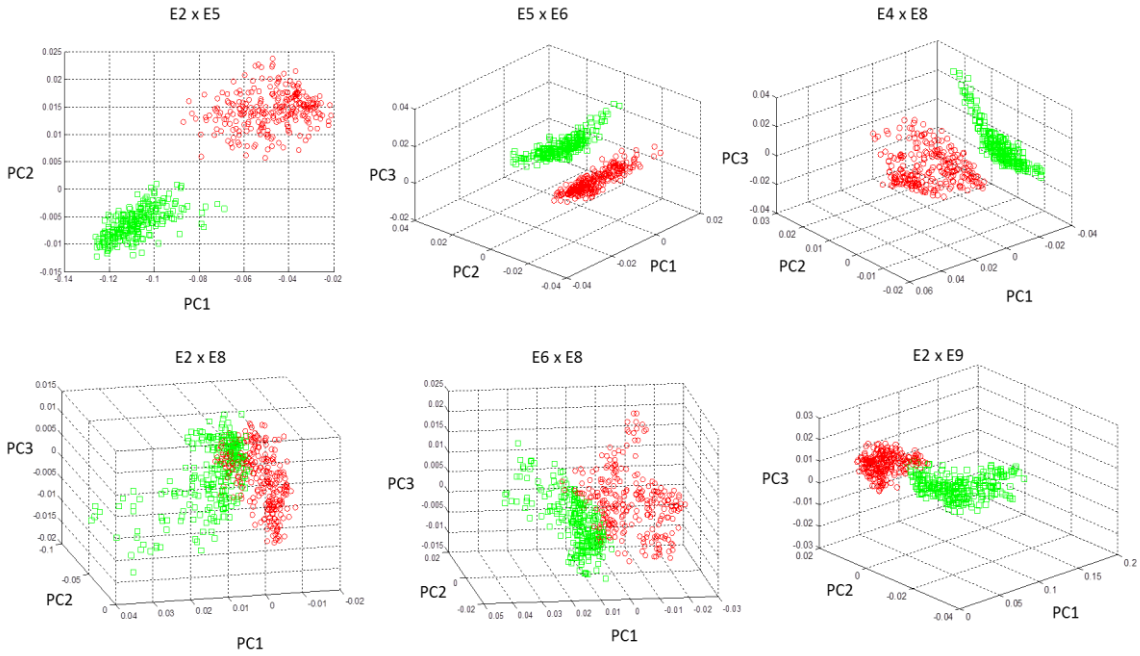


Figura 23: Gráfico dos escores de PCA para seis combinações de canetas esferográficas com tinta a base de óleo. Na primeira linha são mostrados os gráficos de escores nos quais a PCA obteve sucesso na discriminação. Na segunda linha estão exemplos dos casos nos quais a PCA não foi eficaz.

Os três primeiros gráficos de escores da Figura 23 exemplificam os casos em que a PCA foi eficiente para reconhecer a presença de dois grupos produzidos com canetas diferentes. Nesses casos em que houve discriminação, a técnica PP mostrou resultados análogos aos da PCA usando poucas componentes, cerca de 4 a 7 PC's foram suficientes. A segunda linha da Figura 23 mostra três gráficos de escores da PCA que exemplificam os resultados em que apenas PP foi capaz de discriminar entre os dois tipos de tintas comparadas. Na Figura 24 são apresentados os gráfico dos escores de PP para os três exemplos anteriores em que a PCA não obteve sucesso. Pode-se observar que não foram necessárias muitas PC's para obter uma boa discriminação entre os dois grupos nesses casos.

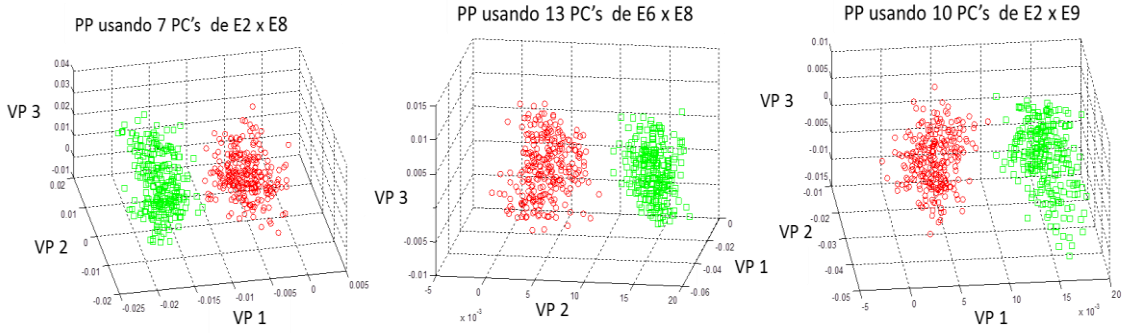


Figura 24: Gráfico dos escores das análises de PP para três combinações de canetas. Onde VP é sigla para Vetor de Projeção.

A combinação das canetas E2 e E6 foi a mais problemática entre as esferográficas, mesmo usando um número elevado de PC's para a análise de PP não foi possível discriminar as duas amostras. As Figuras 25a e 25b mostram os gráficos de escores da PCA e de PP para esse caso.

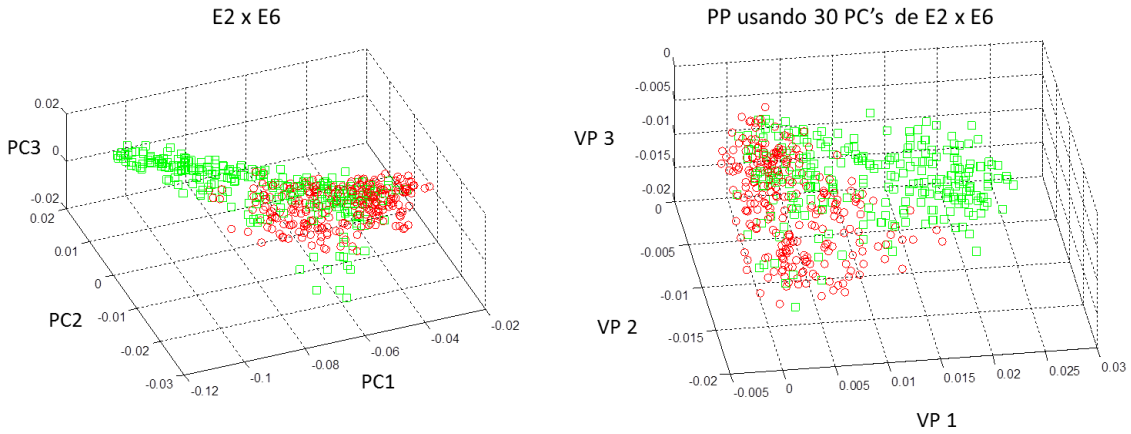


Figura 25: Gráfico dos escores das análises de (a) PCA e (b) Project Pursuit para a combinação entre E2 e E6.

As canetas do tipo Gel foram mais facilmente discriminadas entre si que as canetas esferográficas. Para as 3 combinações feitas, a PCA obteve sucesso na discriminação das tintas e PP foi aplicada apenas para reforçar os resultados. A Figura 26 mostra que as combinações da tinta da caneta G1 com G2 e G3 foram discriminadas por PCA nas duas primeiras PC's. Já para a comparação entre G2 e G3, as diferenças entre as tintas só foi evidenciada na quarta PC.

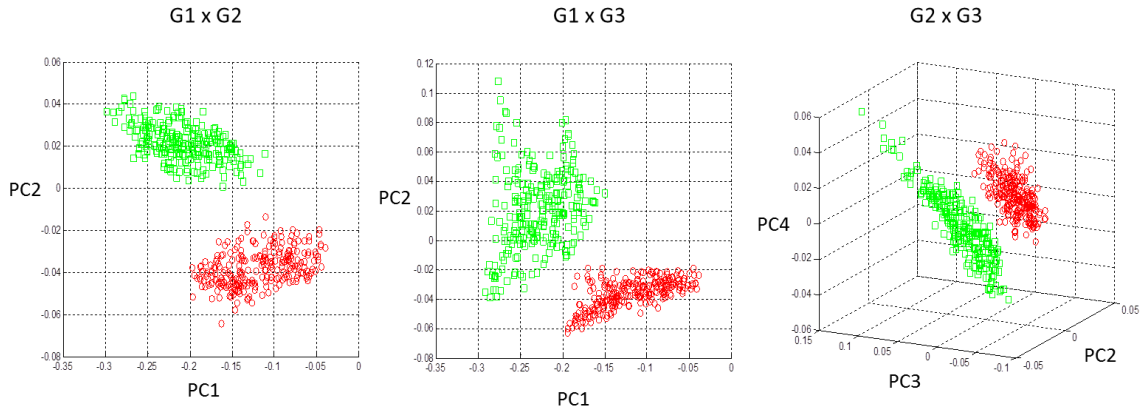


Figura 26: Gráfico dos escores da PCA para as combinações de canetas do tipo Gel.

As combinações de canetas com tinta hidrográfica apresentaram discriminação para cinco dos seis pares de caneta usando apenas PCA. No entanto, para a combinação entre as canetas

R1 e R2 não foi obtida discriminação com PCA. A análise PP também não foi capaz de encontrar informações que diferenciem as duas canetas.

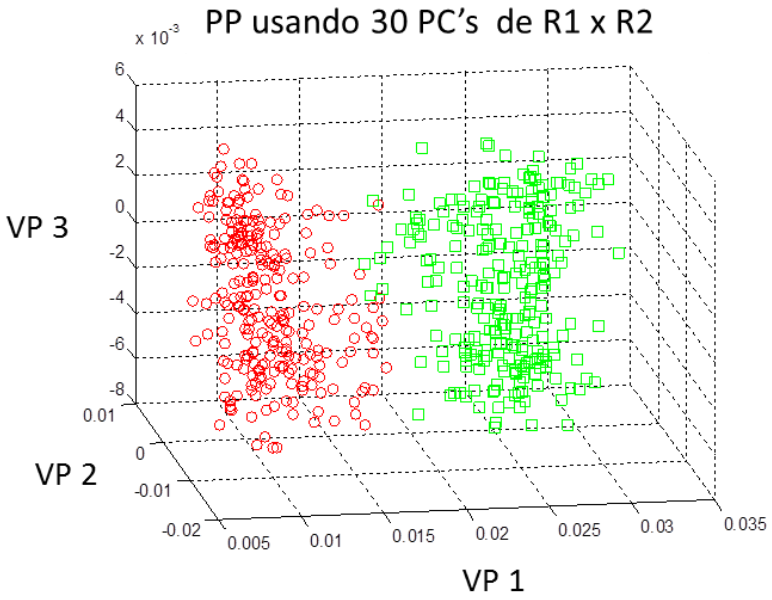


Figura 27: Gráfico dos escores da análise de Project Pursuit para a combinação entre as canetas R1 e R2.

Embora o gráfico de escores de PP mostre uma tendência de discriminação entre as duas canetas, a separação em os dois grupos não clara. As informações obtidas para cada caneta parecem bastante semelhantes, dessa forma a discriminação obtida usando mais que 30 PC's não deve estar correlacionada com as informações químicas, como discutido no capítulo 1.

Os resultados obtidos para as combinações de canetas com diferentes tipos de tinta estão resumidos no APÊNDICE B, totalizando 75 pares de canetas diferentes. Para os 27 pares resultantes da combinação de canetas esferográficas (E1-9) e do tipo Gel (G1-3), a PCA obteve sucesso na discriminação de 21 pares e PP foi capaz de discriminar todos os pares. As combinações contendo a caneta G2 não apresentaram dificuldade na discriminação por PCA e nem por PP. Para os seis pares que a PCA não obteve sucesso, a análise de PP precisou de até 20 PC's para discriminá-las.

A combinação das canetas esferográficas e hidrográficas (H1,2 e R1,2) resultou em 36 pares distintos, dos quais 31 foram corretamente discriminados usando apenas PCA. PP obteve sucesso na discriminação de 35 pares. Apenas a combinação das canetas E2 e R2 apresentou problemas na diferenciação. Como pode ser visto na Figura 28, os escores de PP usando 28 PC's para o par E2 *x* R2 apresenta uma separação razoável das duas canetas,

entretanto, também é possível identificar uma separação dentro de cada conjunto de dados de cada caneta. Isso implica dizer que a discriminação dos dois grupos só foi obtida em uma situação extrema, em que as informações utilizadas permitem diferenciar até amostras do mesmo grupo. Dessa forma a discriminação não foi obtida de forma satisfatória por PP. As outras 5 combinações de canetas esferográfica e hidrográficas que foram discriminadas apenas por PP utilizaram um número reduzido de PC's, variando de 4 a 8 componentes.

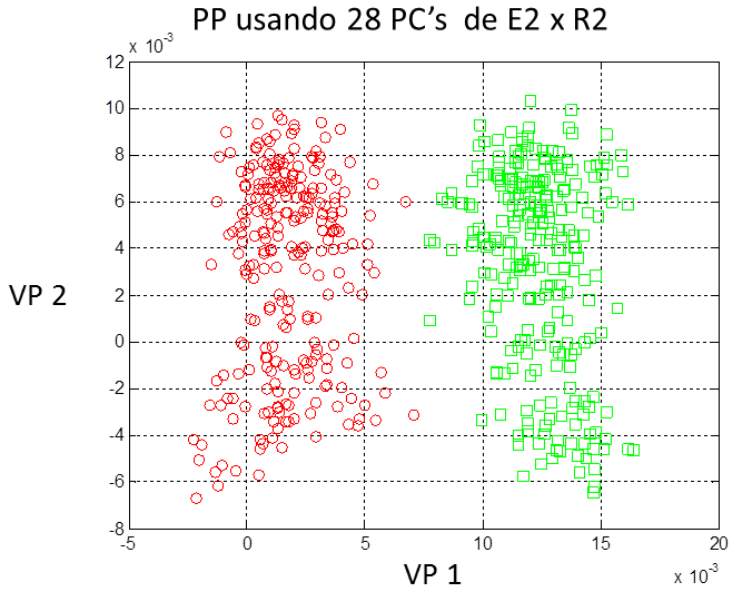


Figura 28: Gráfico dos escores da análise de Project Pursuit para a combinação entre as canetas E2 e R2.

Os últimos 12 pares de combinações obtidos com as canetas do tipo Gel e Hidrográficas foram discriminados corretamente quando a análise de PP foi aplicada usando até 12 PC's. Para 4 combinações, a PCA não obteve sucesso na discriminação das tintas.

6 CONCLUSÃO – DISCRIMINAÇÃO DE TINTAS

Foi estabelecido que o melhor pré-processamento para os espectros de infravermelho médio obtidos através de imagem por ATR deve constar de correção de ruído e de espalhamento da radiação. De acordo com os resultado apresentados, a 1ª derivada com filtro de suavização S-G, janela de 9 pontos e polinômio de 2ª ordem foi selecionada para corrigir os efeitos indesejáveis e maximizar a discriminação entre canetas de tinta preta.

Observando-se os resultados, foi possível definir que PP tem maior capacidade discriminante para tintas preta de canetas em comparação com a PCA. A análise dos escores da PCA permitiu discriminar corretamente 89 pares de canetas, cerca de 74,2% das combinações. Por outro lado, PP obteve sucesso na discriminação de 117 pares, ou seja, em 97,5 % das combinações testadas.

7 SOLUÇÃO DE TESTE-CEGO

Depois de fazer uma avaliação visual minuciosa em cada amostra preparada para o teste-cego, foram levantadas hipóteses sobre como possíveis falsificações teriam sido forjadas em cada caso. Baseado nessas suspeitas, foram definidos o número e a posição de cada amostragem usando HSI-MIR.

Para exemplificar os passos seguidos durante a amostragem dos pontos em cada registro, as aquisições feitas nas amostras TCA1, TCC1 e TCF1 serão detalhadas. Todas as amostras preparada com as indicações dos pontos de amostragem estão reunidas no APÊNDICE C. Como pode ser visto na Figura 29, a amostra TCC1 contém um número 8, no entanto, existe a possibilidade de se tratar do número 3 ou 5 convertidos em 8. Para testar essas hipóteses foram feitas 4 amostragens. Se **a** e **b**, forem semelhantes entre si e diferentes de **c**, o registro original é o número 3. Se **b** for similar a **c** e diferente de **a**, trata-se do número 5. Se forem todos iguais o registro é de fato o número 8.

Figura 29: Amostra TCC1 com detalhamentos dos pontos de amostragem das imagens hiperespectrais no MIR.

A Figura 30 apresenta a amostra TCA1 do teste-cego, que contém o número **4**. Existe a possibilidade de que esse registro seja o número **1,** convertido em número **4**. Foram necessários dois pontos de amostragem, além da amostragem do papel puro, para testar essa hipótese. Se o conjunto de dados da HSI-MIR do ponto **a** for diferente do conjunto obtido no ponto **b**, trata-se de uma adulteração. Se os conjuntos forem similares, o registro é o número 4.

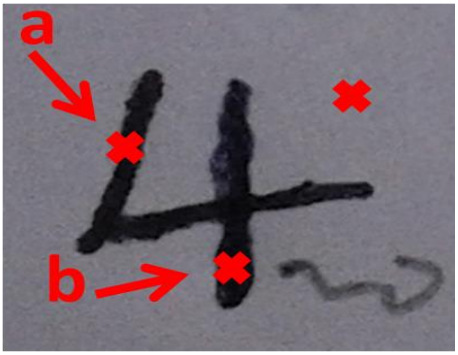


Figura 30: Amostra TCA1 com detalhamentos dos pontos de amostragem das imagens hiperespectrais no MIR.

Para amostras semelhantes à TCF1 (Figura 31), a escolha dos pontos de aquisição é mais simples. Como se trata de um caso em que a única forma de falsificação é por adição de texto, basta fazer amostragens em cada número de acordo com a Figura 31. Aparentemente se trata do número **100**, porém, existe a possibilidade de ser o número **10** com adição de um zero (se **a** e **b** forem similares entre si e diferentes de **c**) ou ainda de ser o número **1** com adição de dois zeros (se **a** for diferente de **b** e **c**).

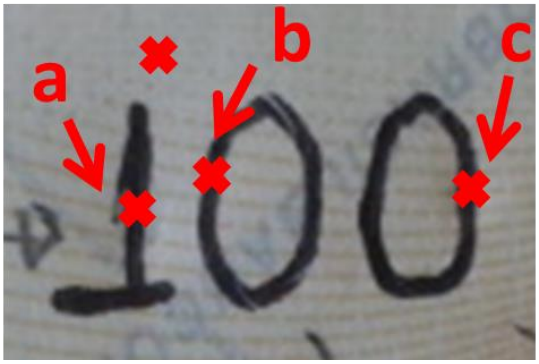


Figura 31: Amostra TCF1 com detalhamento dos pontos de amostragem das imagens hiperespectrais no MIR.

Para todas as amostras foram selecionados pontos para amostragem do papel. No caso do papel de cheque a amostragem foi realizada em uma região próxima do número e com as mesmas características de cor e textura. Essas imagens do papel foram usadas para realizar a extração da ROI. Um aspecto negativo da técnica de imagem por ATR é que o cristal deforma a superfície do papel.

7.1 Pré-Processamento de Dados HSI-MIR

O pré-processamento espectral das imagens obtidas para o teste-cego no MIR foi realizado de acordo com o descrito para o melhor pré-processamento no item **5.1.** Para cada imagem obtida, a matriz de dados foi desdobrada e submetida 1ª derivada com filtro de suavização SG, janela de 9 pontos e polinômio de 2ª ordem.

Uma vez que as amostras do teste-cego simulam amostras reais, efeitos da grafia do colaborador são possíveis interferentes para a aquisição das imagens. Em alguns pontos de amostragem as linhas de tinta eram estreitas, curtas ou com falhas, dessa forma a imagem obtida poderia representar a tinta e também uma pequena parte do papel. Portanto, além do pré-processamento espectral, as imagens hiperespectrais no MIR precisaram ser submetidas à seleção de pixels para maximizar as informações relacionadas com a tinta e reduzir a influência do papel.

A seleção de ROI foi realizada usando os histogramas de frequência da PCA. Para cada amostra foram remontadas imagens com os dados pré-processados da tinta e do papel. Essas imagens foram concatenadas uma acima da outra e em seguida a extração da ROI foi aplicada. Os limites nos histogramas foram escolhidos de modo que todos os pixels da matriz de papel fossem excluídos, de modo a restar apenas os pixels com mais informação da tinta. As Figuras 32a-e mostram, como exemplo, os passos da seleção de pixel para um ponto da amostra TCF1. A partir da imagem dos escores para 1ª PC (Figura 32a) foi escolhida a faixa a ser tomada no gráfico dos histogramas de frequência (Figura 32c) que resultou na seleção dos pixels marcados em vermelho (Figura 32b). As matrizes de pixels ou espectros obtidas para todos os pontos de cada amostra foram então submetidas à PCA e PP.

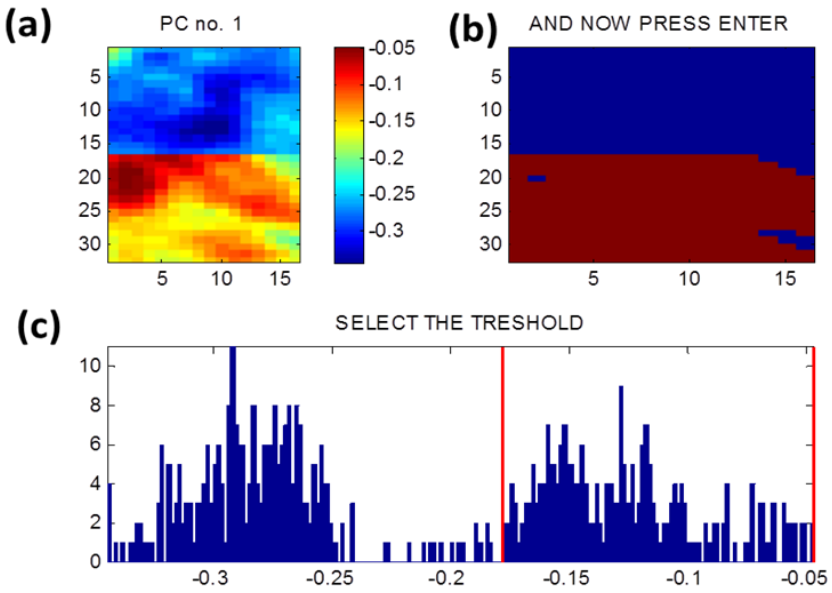


Figura 32: (a) Imagem dos escores da PC1 da matriz do papel sobre a matriz da tinta; (b) imagem dos pixels da tinta (vermelho) depois de removidos os pixels com informação apenas do papel; (c) gráfico dos histogramas de frequência dos pixels para PC1.

7.2 Pré-Processamento de Dados HSI-NIR

As imagens hiperespectrais no NIR foram adquiridas na faixa espectral 928-2524 nm, mas em decorrência de excesso de ruído e anomalias da medida os 20 a 30 primeiros números de onda e os 6 a 16 últimos números de onda foram removidos dos espectros, resultando em uma faixa de trabalho de 1111 e 2425 cm^{-1}. As Figuras 33a e b exemplificam as remoções efetivadas nos espectros de todas as HSI-NIR's. O número de variáveis removidas dependeu das características espectrais de cada imagem. No exemplo da Figura 33a foram removidas as 30 primeiras e 6 últimas variáveis resultando no conjunto de espectros da Figura 33b.

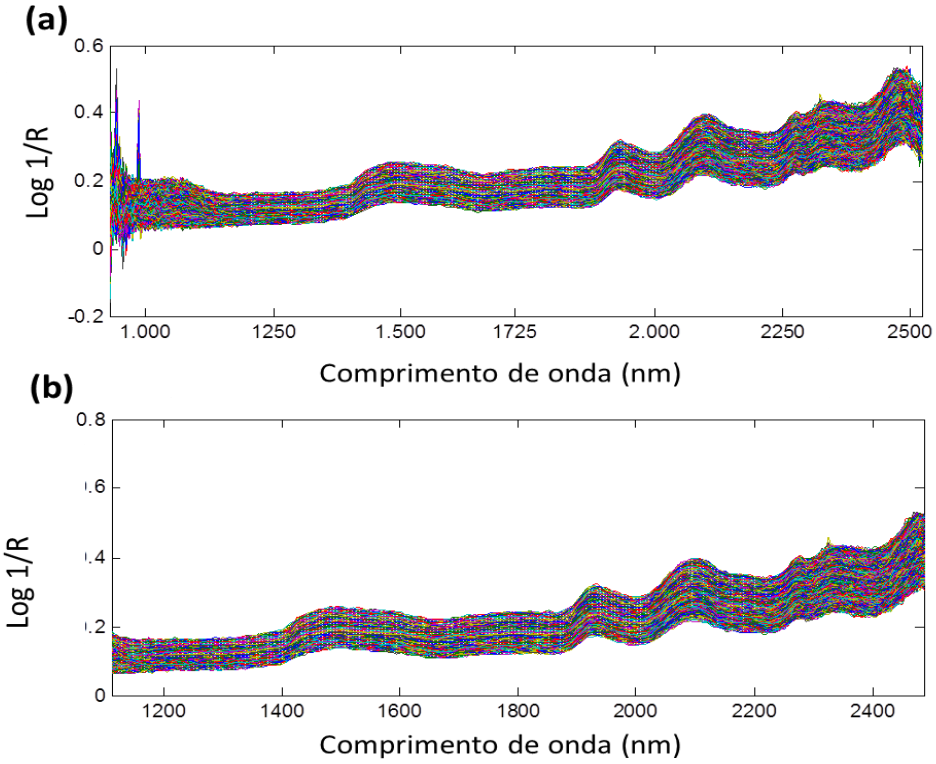


Figura 33: (a) Espectros brutos e (b) cortados da amostra TCF4.

As técnicas de pré-processamento espectral foram aplicadas inicialmente em uma amostra aleatória para detectar o máximo de discriminação entre a tinta da caneta e o papel usado. Como a composição dos números escritos não era conhecida pelo avaliador, buscou-se identificar em que condições de pré-processamento as amostras apresentavam todo ou parte dos números escritos com mais clareza nas imagens dos escores das PC's. A imagem TCF2 será usada para demonstrar como foi obtido o melhor pré-processamento.

O primeiro passo foi avaliar o impacto do pré-processamento na correção dos efeitos indesejáveis resultantes da técnica de aquisição e dos aspectos físicos do papel e em seguida as imagens dos escores foram comparadas. Como pode ser visto na Figura 33, o maior problema no espectro está relacionado com o espalhamento dos espectros (efeito aditivo). Para correção desse efeito foram comparados os espectros tratados com SNV, com 1ª e 2ª derivadas com filtro SG, polinômio de 2ª ordem e janela de 7 e 11 pontos respectivamente. O pré-processamento usando SNV apresentou resultados muito melhores na correção do efeito aditivo que os demais (Figura 34a-c) e também obteve boa discriminação da tinta e do papel, além de corrigir os efeitos de compactação do papel pelo cristal de ATR, como se observa na imagem dos escores da PC1 (Figura 35a-c). Dessa forma a técnica SNV foi utilizada para tratar as demais HSI-NIR's.

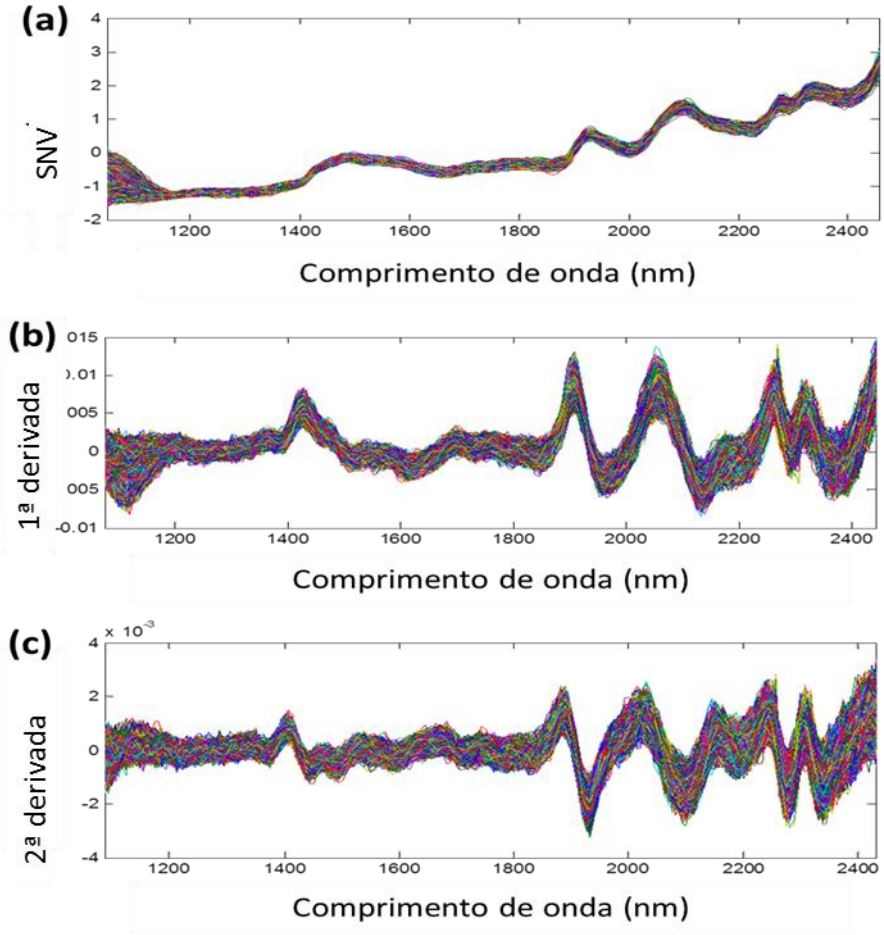


Figura 34: Espectros pré-processados com (a) SNV; (b) 1ª derivada SG, janela de 7 pontos e polinômio de 2ª ordem; (c) 2ª derivada SG, janela de 11 pontos e polinômio de 2ª ordem.

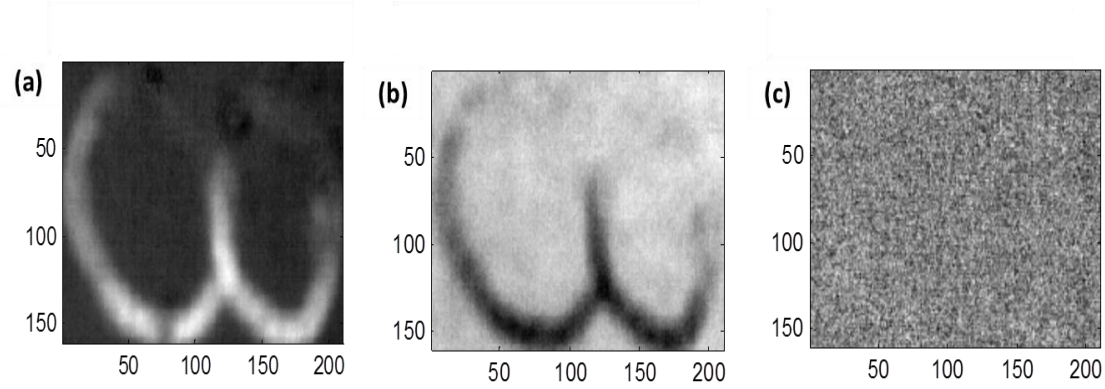


Figura 35: Imagens dos escores da PC1 par os espectros pré-processados com (a) SNV ; (b) 1ª derivada SG, janela de 7 pontos e polinômio de 2ª ordem; (c) 2ª derivada SG, janela de 11 pontos e polinômio de 2ª ordem.

Em um trabalho anterior com as mesmas canetas, Silva e colaboradores (2014a) chegaram a mesma conclusão quanto ao melhor pré-processamento para as imagens

hiperespectrais de tintas de caneta na região do NIR, mas as amostras e os objetivos são diferentes, dessa forma foi importante confirmar as observações anteriores.

7.3 PCA e PP Associadas à HSI-NIR e HSI-MIR para Resolução de Testes-Cego

Apenas alguns dos testes-cego serão detalhados usando todos os resultados obtidos e que levaram a conclusão em cada caso. Para as demais amostras os resultados serão apenas descritos de acordo com a necessidade para chegar à conclusão. Para os pixels obtidos das imagens no MIR, os gráficos de escores da PCA e de PP sempre apresentarão os mesmos símbolos e cores para os conjuntos amostrados em cada ponto. Os pixels amostrados no ponto **a** serão simbolizados pelo asterisco preto (*), para o ponto **b** será utilizado o triângulo verde (Δ) e no ponto **c** os pixels serão simbolizados por círculos vermelhos (o).

Amostra TCC1

A primeira amostra avaliada foi a TCC1 que continha o número **8** escrito em pedaço de folha de papel. Foram adquiridas HSI-MIR em apenas três pontos do número como detalhado na Figura 29. A PCA e análise de PP indicaram que os três pontos amostrados (a, b e c) são distintos entre si, o que não permite inferir sobre a possível adulteração (Figura 36 a e b). No entanto, as imagens dos escores da PCA e de PP para a imagem no NIR evidencia claramente a forma como o número foi adulterado e o motivo que levou ao resultado curioso usando HSI-MIR.

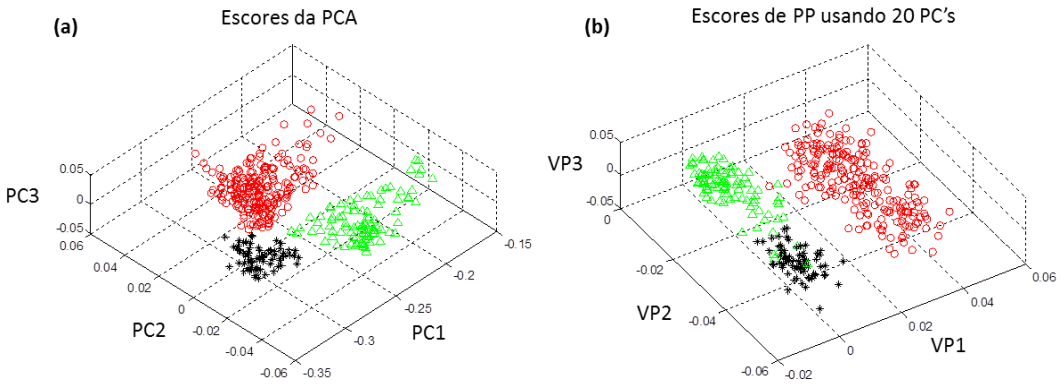


Figura 36: Gráfico dos escores da (a) PCA e de (b) PP usando 20 PC's para a mostra TCC1.

Avaliando as imagens dos escores da PCA e de PP para os dados no infravermelho próximo foi possível concluir que o texto original era o número **5** (Figura 37 a e b) que foi convertido em **8**. O número **5** foi totalmente recoberto com uma tinta diferente, dessa forma a amostragem no ponto **a** feita com o acessório de imagem por ATR abrangeu uma área que

continha as duas tintas separadas como pode ser visto na imagem dos escores da Figura 37a. Por ter informações sobre as duas canetas, o conjunto de dados **a** pode ter ficado igualmente dividido entre os pixels dos outros dois pontos. Os pontos circulares presentes na imagem dos escores da PC1 e do VP1 são resíduos deixados pela deformação da superfície do papel causada pela amostragem com o acessório de ATR. Esse resultado está em consonância com a revelação do teste-cego pelo colaborador.

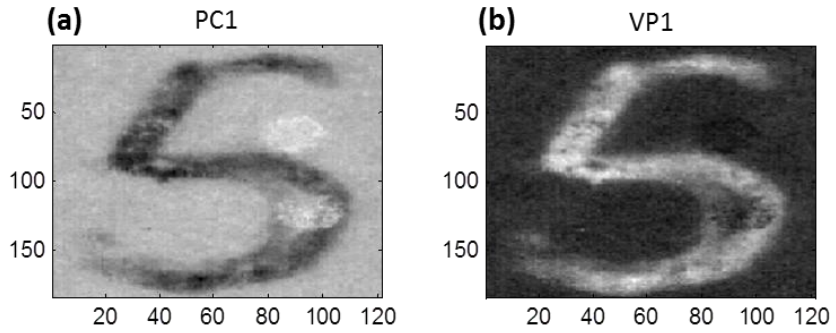


Figura 37: Imagem dos escores da (a) PC1 e de (b) PP usando 6PC's para a mostra TCC1.

Amostra TCC2

O número **9** contido na amostra TCC2 aparenta ter sido produzido a partir do número **1** mediante adição de texto. Foram obtidas duas imagens na região do MIR nos pontos indicados na Figura 38. Como mostrado na Figura 39, a PCA da HSI-MIR foi suficiente para discriminar os dois conjuntos de dados com facilidade, indicando ter havido adulteração do número original. As análises de PP e PCA da HSI-NIR não colaboraram para o resultado, uma vez que não foi possível distinguir as tintas do papel para esse caso. De acordo com o colaborador **C**, o número original era o **4** convertido em **9** por adição da parte superior, ou seja, a discriminação feita mostrou-se correta, embora a hipótese inicialmente levantada (número 1 convertido em 9) não foi correspondeu a fraude simulada.

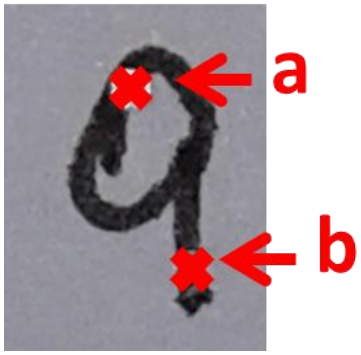


Figura 38: Amostra TCC2 indicando os pontos de amostragem usando HSI-MIR.

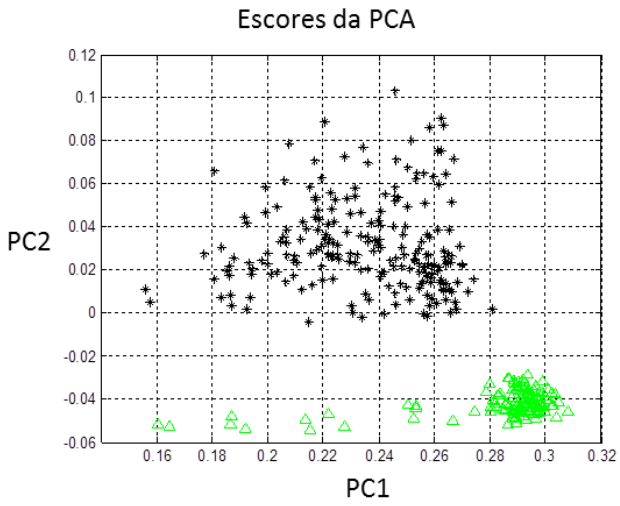


Figura 39: Gráfico dos escores da PCA para amostra TCC2.

Amostra TCC3

Para as imagens no infravermelho médio da amostra TCC3, a remoção dos pixels mais relacionados com o papel reduziu a matriz de dados a menos de 100 pixels e por esse motivo a avaliação por PCA e PP não foi considerada representativa. A análise das imagens no NIR também não puderam discriminar a tinta e o papel. Não foi possível chegar a uma conclusão sobre esse teste.

Amostra TCC4

Para a amostra TCC4 nenhum dos pontos amostrados usando HSI-MIR foi discriminado dos demais usando PCA ou PP, dessa forma o avaliador concluiu que o número **8** deve ter sido produzido com uma única caneta (Figura 40 a e b). Os resultados da PCA e de PP para a imagem do número no NIR corroboram com esse resultado, uma vez que a PC2 e PC3 evidenciam o número **8**, assim como os vetores de projeção 2, 3 e 4 (Figura 41a e b).

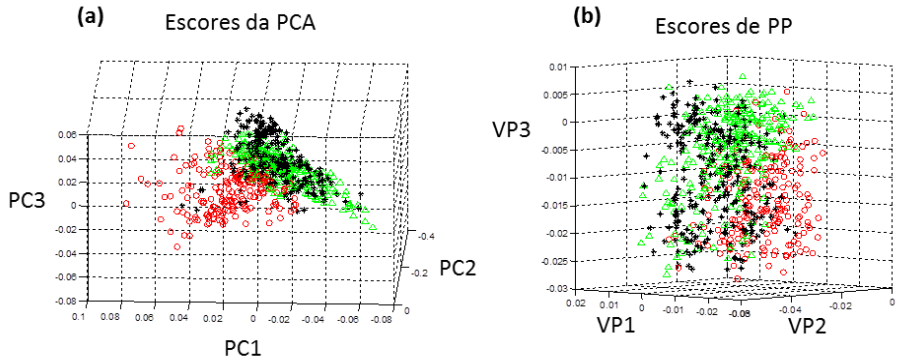


Figura 40: Gráfico dos escores (a) da PCA e de (b) PP usando 35 PC's para amostra TCC4.

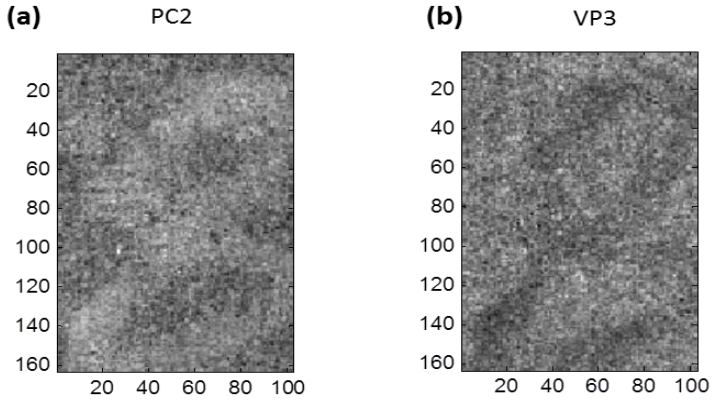


Figura 41: Imagem dos escores da (a) PCA e de (b) PP usando 6PC's para a amostra TCC4.

Amostra TCC5

Foram feitas duas aquisições de HSI-MIR no número **10** contido na amostra TCC5: uma amostragem no número **1** (um) e outra no **0** (zero). Como não houve discriminação entre os dois conjuntos, concluiu-se que não se trata de uma falsificação. As imagens dos escores da PCA e de PP das HSI-NIR's mostram o perfil do número **10** completo, portanto o resultado corrobora com o anterior. A conclusão está correta de acordo com o colaborador.

Amostra TCC6

A amostragem usando MIR em TCC6 foi realizada nos dois algarismos que formam o número **10.** Os resultados da PCA e de PP não permitem distinguir os dois conjuntos de dados, levando a conclusão de que se trata de amostra real. No entanto, as imagens dos escores da PCA e PP para as imagens no NIR mostram claramente que o número **0** (zero) foi escrito com uma caneta de diferente da usada para escrever o número **1**(um)(Figura 42)**.**Foi selecionada apenas uma área uma parte da imagem no NIR devido a forte influência da área deformada pelo cristal de ATR na parte superior do algarismo **1**.

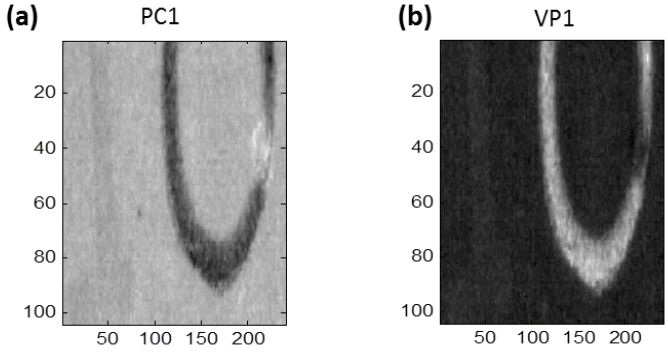


Figura 42: Imagem dos escores da (a) PCA e de (b) PP usando 6PC's para a amostra TCC6.

Amostra TCA1

A primeira amostra produzida pelo colaborador A pode ser vista na Figura 30 com os pontos de amostragem de HSI-MIR apontados. Os gráficos de escores da PCA e de PP apresentaram uma separação discreta entre os conjuntos de dados adquiridos nos pontos **a** e **b**, no entanto poucos pixels da amostragem feita em **a** foram discriminados do papel e o resultado pode ter sido comprometido. Avaliando as imagens dos escores da PCA e PP para a imagem no NIR fica evidente que o número **4** escrito na amostra é uma adulteração feita no algarismo **1** (Figura 43). Essa solução satisfaz exatamente as condições de preparo da amostra segundo o colaborador A.

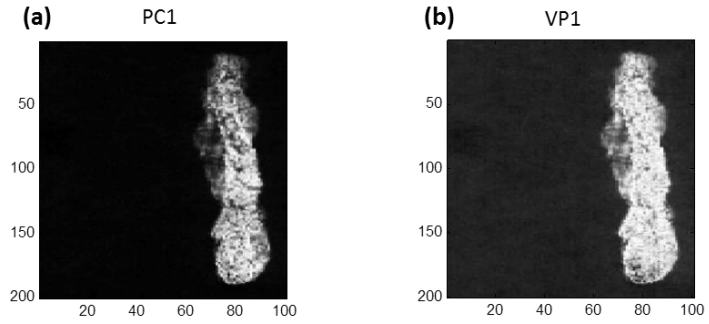


Figura 43: Imagem dos escores da (a) PCA e de (b) PP usando 6PC's para a amostra TCA1.

Amostra TCA2

Para o teste-cego TCA2 não foi obtida uma boa discriminação entre as três imagens obtidas no MIR. Apenas um ponto se diferenciou um pouco dos demais, no entanto foi insuficiente para determinar se houve adulteração do número. Quando PCA foi aplicada a imagem obtida no NIR, tornou-se evidente a conversão do número **1** em **7.** Na Figura 44b é

possível identificar os traços utilizados para fazer a adulteração do registro. Os resultados de PP corroboram com a PCA e levam a conclusão correta segundo o colaborador A.

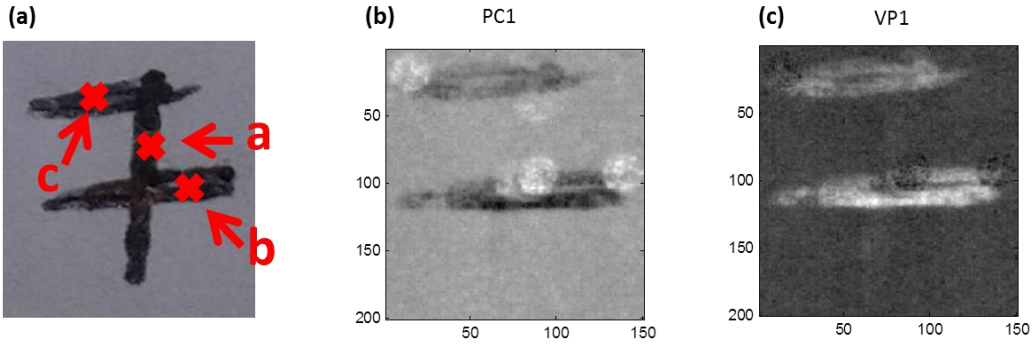


Figura 44: (a) Amostra TCA2, (b) imagem dos escores da PCA e de (b) PP usando 6PC's.

Amostra TCA3

A PCA e análise de PP para o conjunto de dados da amostra TCA3 revelou que foram usadas duas canetas diferentes para produzi-la. O conjunto de dados obtidos para o ponto **b** foi discriminado dos demais. Dessa forma o registro original é o número **1** que foi convertido em **7** (ver Figura 45a). A análise da HSI-NIR para a amostra TCC3 não permitiu inferir sobre a autenticidade do registro produzido, uma vez que as diferenças entre a tinta e o papel não foram detectadas.

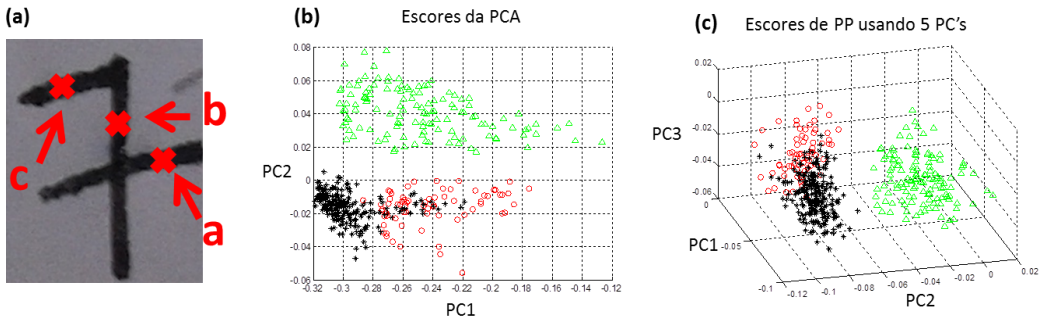


Figura 45: (a) Amostra TCA3, (b) gráfico dos escores da PCA e de (b) PP usando 5PC's.

Amostra TCA4

No teste-cego TCA4, as técnicas PCA e PP não identificaram grupos diferentes para as amostragens feitas em três pontos diferentes do número usando MIR. As imagens dos escores da PCA e de PP nos dados de NIR revelaram a presença do número **8** na amostra sem nenhum indício de falsificação. Embora não tenham sido encontradas evidencias de adulterações no algarismo, o colaborador revelou que a amostra continha o número **2** convertido em número **8** usando uma caneta diferente.

Amostra TCA5

Os conjuntos de dados das imagens no MIR adquiridas em dois pontos do número **4** da amostra TCA5 foram discriminados com facilidade usando PCA e/ou PP. As imagens dos escores da PCA e de PP para a imagem no NIR corroboram com esse resultado, uma vez que foi detectada apenas uma parte do número. Com base no exposto na Figura 46a-d, o registro original era o número **1** (um) que foi transformado em **4** usando outra caneta. Esse resultado está de acordo com o revelado pelo colaborador A.

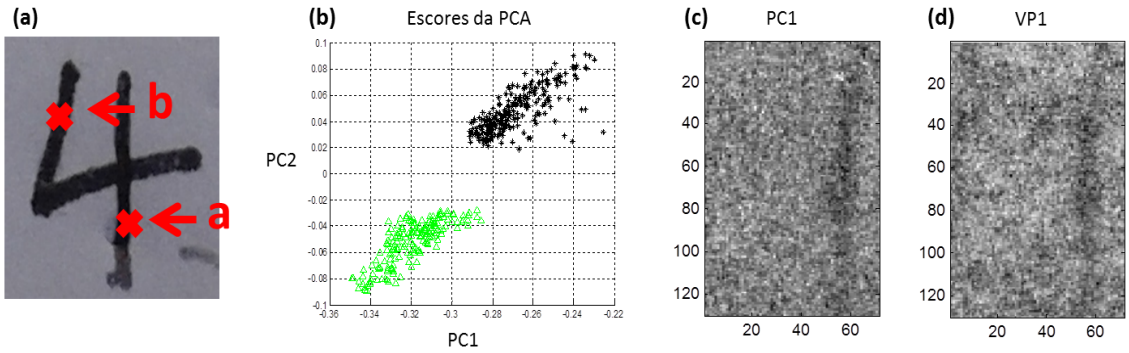


Figura 46: (a) Amostra TCA5, (b) gráfico dos escores da PCA (MIR), (c) imagem dos escores da PCA e (d) imagem dos escores de PP usando 6PC's (NIR).

Amostra TCA6

Foram adquiridas três imagens no MIR de parte das linhas que formam o algarismo **8** da amostra TCA6**.** Os conjuntos de dados não foram discriminados usando PCA e PP. Quando essas técnicas foram aplicadas a imagem adquirida no NIR os resultados foram bem diferentes, mostrando claramente o algarismo original que foi convertido em **8.** As Figuras 47a-c mostram os resultados obtidos com PCA e PP, onde o número **3** toma forma e se destaca nas imagens dos escores. A PC2 mostra também parte da tinta que foi usada para adulterar o número original. A revelação do teste-cego está de acordo com os resultados obtidos.

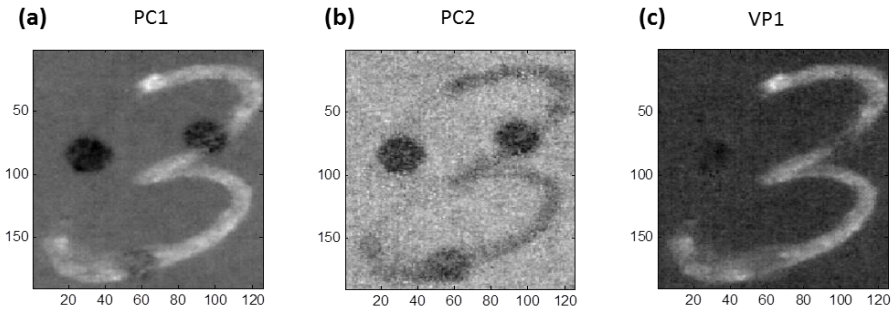


Figura 47: Imagem dos escores da (a) PC1, da (b) PC2 e de (c) PP usando 6PC's.

Amostra TCA7

A amostra TCA7 é semelhante à amostra anterior e tem um resultado similar. Para as imagens no MIR apenas PP usando 27 PC's foi capaz de discriminar uma amostragem das demais. Esse resultado permitiu determinar que a amostra em questão é uma falsificação em que o algarismo **3** foi convertido em **8.** As imagens dos escores da PCA e PP para a imagem no NIR corroboram com esse resultado, pois evidenciam a presença do número **3** na imagem.

Amostra TCA8

Para a amostra TCA8 foram realizadas aquisições em três pontos do número **7** usando o acessório de imagem no infravermelho médio. Entretanto, a amostragem feita na barra horizontal superior do **7** teve muitos pixels removidos e ela foi removida antes de tratar com a PCA e PP por não ser considerada representativa da imagem. A PCA nos dados do MIR e na imagem do NIR foi suficiente para revelar a fraude, em que o número **1** foi convertido em **7** usando outra caneta. De acordo com o colaborador A essa conclusão está correta.

Amostra TCA9

As imagens adquiridas no MIR em dois pontos do número **4,** registrado na amostra TCA9, foram separadas facilmente usando apenas duas PC´s. Assim é possível afirmar que houve adulteração do número **1** com uma caneta diferente para formar o número **4** visualizado na amostra. A imagem dos escores da PCA da imagem amostrada no NIR destaca nas primeiras PC's a parte que foi adicionada usando esta outra caneta, confirmando assim o resultado obtido com a análise dos dados das imagens no MIR.

Amostra TCA10

Para a amostra TCA10 não foi possível discriminar a tinta usada para escrever o número **8** do papel onde foi escrito usando, PCA e PP na HSI-NIR. Porém, os conjuntos de dados obtidos na região do MIR foram suficientes para determinar autenticidade do registro feito. Como pode ser visto na Figura 48a-c, os conjuntos de dados **a** e **b** foram separados do conjunto **c** pela PC2. O gráfico de escores da análise de PP usando muitas PC's confirma que os conjuntos de pixels **a** e **b** são provenientes de amostras com a mesma tinta. Os resultados estão de acordo com a descrição do colaborador A.

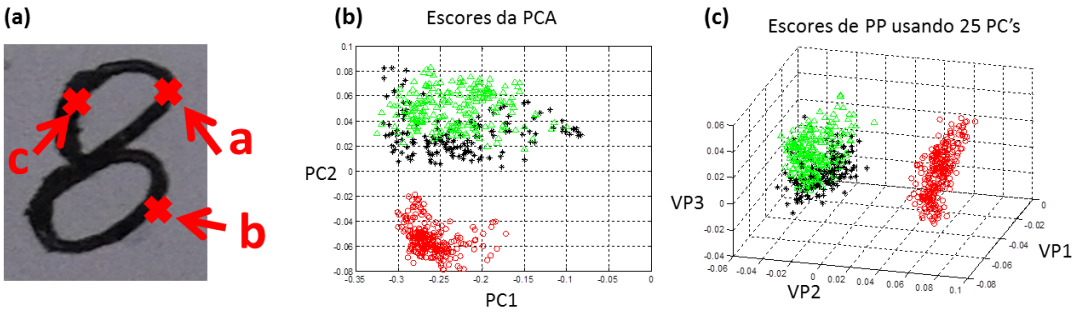


Figura 48: (a) amostra TCA10, gráfico dos escores (b) da PCA e de (c) PP usando 25PC's.

Amostra TCA11

Foram adquiridas imagens no MIR em três pontos do número **4** contido na amostra TCA11. Depois de aplicar PCA e PP nos conjuntos de dados das imagens, foi observada a discriminação entre as duas partes do algarismo analisado, ou seja, a amostra TCA11 é uma falsificação. A análise da imagem no NIR confirmou o resultado obtido usando os dados de MIR. Assim é possível afirmar que o registro original é o número **1** que foi convertido no número **4** usando para tal uma caneta diferente.

Amostra TCA12

O teste-cego TCA12 é similar ao teste anterior, no entanto foram feitas amostragens em apenas duas partes do número **4.** A discriminação entre os dois conjuntos de dados das imagens no MIR foi obtida com apenas duas componentes na PCA, no entanto poucos pixels de uma das duas imagens foram utilizados na análise e isso pode ter comprometido o resultado. Felizmente os resultados da PCA e da análise de PP para a imagem no NIR confirmaram os resultados obtidos anteriormente, e, sendo assim, novamente foi determinado que o texto original é o número **1**.

Amostra TCA13

Aplicando a PCA e PP na imagem no NIR da amostra TCA13, foi possível identificar a presença de duas tintas diferentes na composição do número **7** registrado no papel. Com base nessas informações obtidas é possível concluir que o texto original foi adulterado usando uma segunda caneta para converter o número **1** em **7**. As imagens adquiridas na região do infravermelho médio também foram submetidas à PCA e PP, e mesmo tendo corroborado com os resultados dos dados no NIR, não foram considerados confiáveis por terem sido usados menos de 90 pixels de cada imagem. Segundo o colaborador A, a conclusão sobre este teste está correta.

Amostra TCA14

Na amostra TCA14 foram adquiridas imagens hiperespectrais no MIR em dois pontos (Figura 49a) e em seguida os conjuntos de dados das imagens foram submetidos à PCA e PP. O gráfico de escores da PCA não mostra discriminação clara entre os dois grupos de dados, mas o resultado de Project Pursuit é conclusivo quanto a discriminação das duas amostragens (Figura 49b e c) usando apenas 13 PC's. Para as imagens obtidas na região do infravermelho próximo, a PCA também não obteve sucesso na identificação da falsificação (Figura 49d). A análise de PP também não foi capaz de identificar indícios de fraude nos primeiros vetores de projeção, mas observando atentamente a imagem dos escores do VP4 é possível identificar apenas uma parte do número **4**, ou seja, foram usadas duas canetas para produzir a amostra (Figura 49f). A explicação para as diferenças entre as tintas só estar presente no VP4 pôde ser obtida depois de reveladas as etapas de preparo da amostra. O colaborador A revelou ter primeiro escrito o número **1** e seguida usou outra caneta para adicionar o número **4** exatamente por cima do primeiro registro. Dessa forma as diferenças entre as tintas seriam minimizadas ou até mascaradas.

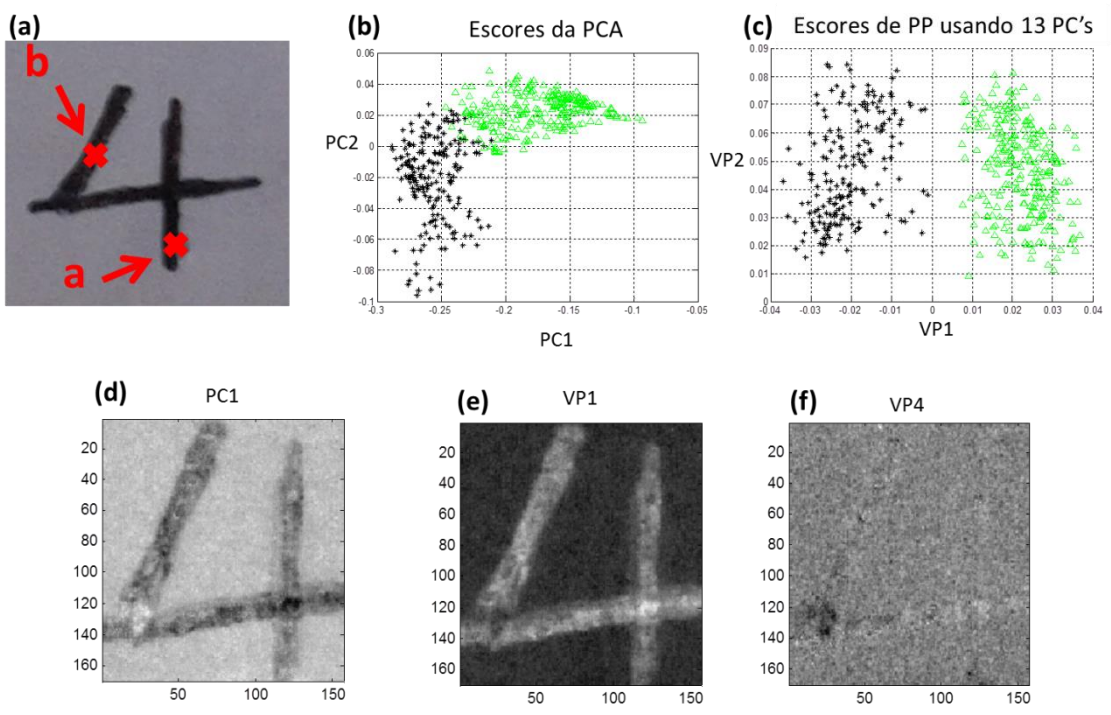


Figura 49: (a) amostra TCA14, gráfico dos escores (b) da PCA e de (c) PP usando 13PC's, (d) imagem dos escores da PCA e (e) Imagem dos escores de PP para VP1 e (f) VP4.

Amostra TCF1

A primeira amostra preparada em papel de cheque apresentava o número **100**, no qual foram realizadas três amostragens: uma aquisição de HSI-MIR em cada algarismo (Figura 50a). Como pode ser visto nas Figuras 50b e c, a PCA permitiu discriminar o conjunto de dados **c** dos conjuntos **a** e **b** e o gráfico de escores de Project Pursuit reforça esse resultado. Baseado nessas observações é possível afirmar que se trata de um caso de adição de texto, em que foi adicionado o algarismo **0** (zero) ao número **10**. A avaliação da imagem no NIR não pode contribuir na conclusão devido às imagens dos escores da PCA e de PP não mostrarem diferença entre as tintas usadas. O colaborador F confirmou esse resultado.

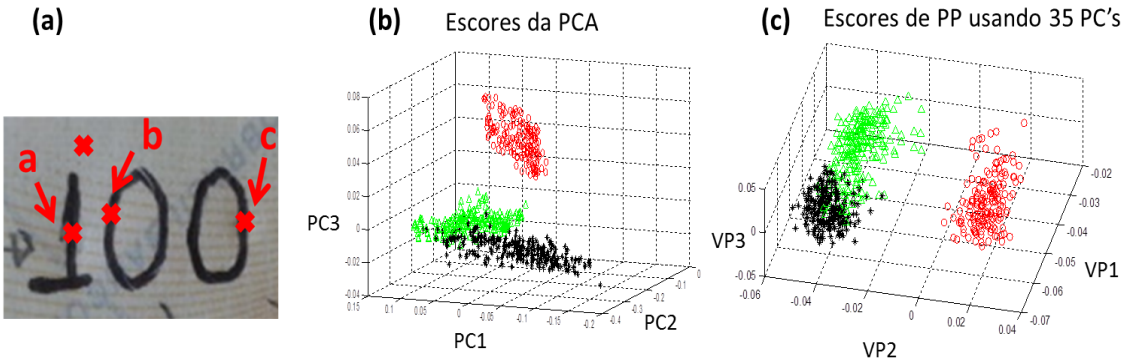


Figura 50: (a) amostra TCF1 e pontos de aquisição, gráfico dos escores (b) da PCA e de (c) PP usando 35PC's.

Amostra TCF 2

Foram adquiridas imagens na região do MIR em três pontos da amostra TCF2 como mostrado na Figura 51a. A PCA não foi capaz de discriminar nenhum dos grupos, apenas a análise de PP foi capaz de evidenciar as diferenças entre **a** e os outros dois conjuntos de dados. Por meio do gráfico de escores de PP dos dados obtidos com HSI-MIR, mostrado na Figura 51b, foi possível definir que o registro original é o número **3.** As imagens dos escores da PCA e de PP da imagem no NIR ajudam a sedimentar essa conclusão.

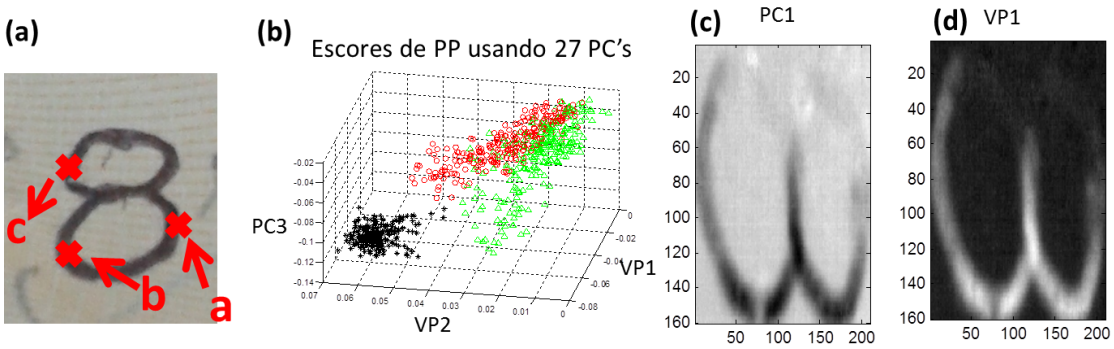


Figura 51: (a) amostra TCF2, (b) gráfico dos escores da PCA e imagem dos escores da (c) PC1 e de (c) PP.

Amostra TCF3

No número **10** contido na amostra TCF3 foram adquiridas duas imagens na região do MIR, uma aquisição em cada algarismo. Os dois conjuntos de dados foram separados usando o mínimo de PC's para a PCA e também para PP. A HSI-NIR submetida à PCA mostra apenas o algarismo **1** nas PC's 3 e 4. PP tem um resultado análogo ao da PCA. Com base nessas informações, chegou-se a conclusão de que os dois números foram escritos com canetas diferentes.

Amostra TCF4

Depois de aplicar PCA na imagem da amostra TCF4, foram identificadas regiões com falhas na aquisição da imagem que causaram a ausência de informação nos pixels. Esses espaços sem informação química da amostra se concentraram principalmente na barra horizontal superiores no número **7** e por esse motivo a amostra essa imagem foi cortada para excluí-los (Figura 52b). A área remanescente continha informação suficiente para solucionar o caso e esta região foi novamente submetida à PCA e PP. Mesmo depois do corte da imagem ainda apareceram alguns pixels anômalos, mas em número reduzido. A imagem dos escores mostra que o número foi produzido com auxílio de duas canetas diferentes, uma para escrever o algarismo **1** e outra para adicionar uma barra e converte-lo em **7** (Figura 52c). As aquisições de imagens na região do MIR foram feitas em três pontos do número **7**, mas, assim como na imagem NIR, só foram considerados como significativas as duas imagens adquiridas na parte inferior do número (Figura 52a). Os gráficos de escores da PCA e de PP mostram a discriminação dos dois conjuntos de dados remanescentes, confirmando o resultado obtido com as técnicas de imagem e processamento anteriores (Figura 52d,e). Estes resultados estão de acordo com as informações sobre a composição do número produzido pelo colaborador.

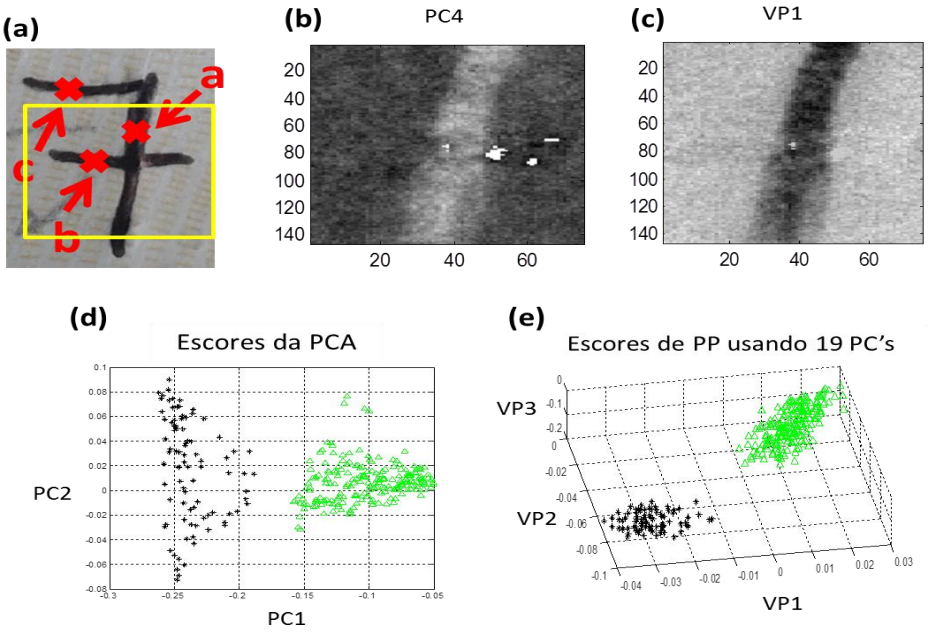


Figura 52: (a) amostra TCF4 indicando pontos de amostragem MIR e área selecionada para avaliação com as técnicas quimiométricas, (b) imagem dos escores da PC4 e do (c) VP1 de PP, (d) gráfico de escores da PCA e (e) gráfico dos escores de PP usando 19 PC's.

Amostra TCF5

Observando as imagens dos escores da PCA e de PP para a imagem no NIR da amostra TCF5, foi possível identificar a adição de um **0** (zero) ao número **10,00** na primeira PC e no quarto vetor de projeção (Figura 53a-c). Os resultados da PCA das imagens no MIR não foram definitivos na discriminação do conjunto de dados **c**, mas a análise de PP mostrou uma clara separação entre **c** e os demais conjuntos de dados usando 27 PC's. Os pixels referentes à vírgula (ponto **d**) se confundiram bastante com os pixels de **a** e **b**.

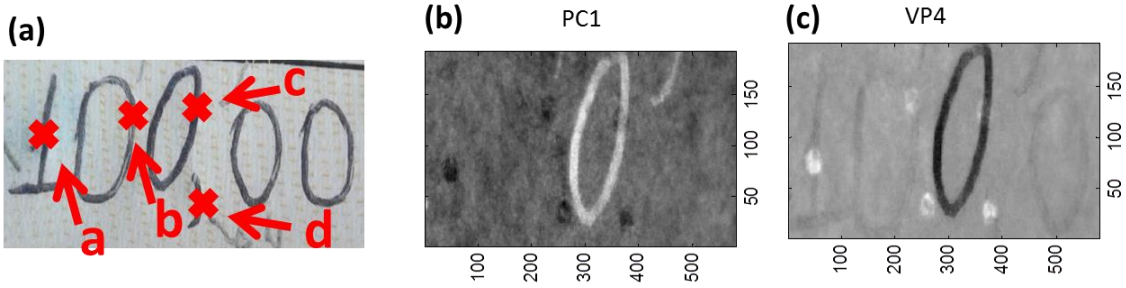


Figura 53: (a) amostra TCF5 com indicações dos pontos amostragem no MIR e imagens dos escores (b) da PC1 e (c) do VP4 de PP.

Amostra TCF6

A falsificação do teste-cego TCF6 foi facilmente identificada usando HSI-MIR e HSI-NIR. Os conjuntos de dados obtidos das imagens no MIR adquiridas em dois pontos do

número **4** foram discriminados já na primeira PC. A imagem dos escores da PCA realizada na HSI-NIR mostrou que foram usadas duas canetas distintas para produzir o número **4**, uma vez que na primeira PC se destaca apenas o algarismo **1** e nas demais PC's aparecem os pixels correlacionas com a tinta usada para adulterar o texto original. O resultado foi confirmado pelo colaborador F.

Amostra TCF7

Foram adquiridas imagens na região do MIR nos dois algarismos que compõem o número **90** da amostra TCF7 e os conjuntos de dados foram submetidos à PCA e PP. Os gráficos de escores para as duas técnicas indicam que os dois conjuntos de dados são diferentes, ou seja, provêm de canetas com tintas diferentes. Os resultados da PCA e de PP da imagem na região do NIR corroboram com a solução encontrada. Segundo as informações dadas pelo colaborador F, a solução está correta.

Amostra TCF8

A amostra TCF8 levantou duas suspeitas: a primeira é que o número **0** foi adicionado usando uma caneta diferente da usada para fazer o número **3** e a segunda suspeita é de que o número três foi recoberto com a mesma caneta usada para adicionar o **0.** Partindo dessa suspeita, foram realizadas as três amostragens indicadas na Figura 54a usando o HSI-MIR. Como resultado da PCA, a parte interna do número **3** foi discriminada da parte externa e do algarismo **0** (Figura 54b), e assim poderia ter havido falsificação. Quando a HSI-NIR foi avaliada com PCA não foi possível diferenciar a tinta e o papel com clareza, apenas a imagem dos escores de PP mostrou com clareza que o número **3 e 0** não apresentam diferenças entre a composição dos dois algarismos que o formam (Figura 54c). O colaborador F confirmou que o número **30** era uma amostra autêntica e que a diferença entre a parte interna e externa do número foi induzida pela sobrecarga de tinta. Assim, o resultado usando HSI-NIR está correto e o resultado usando HSI-MIR representa um falso-positivo.

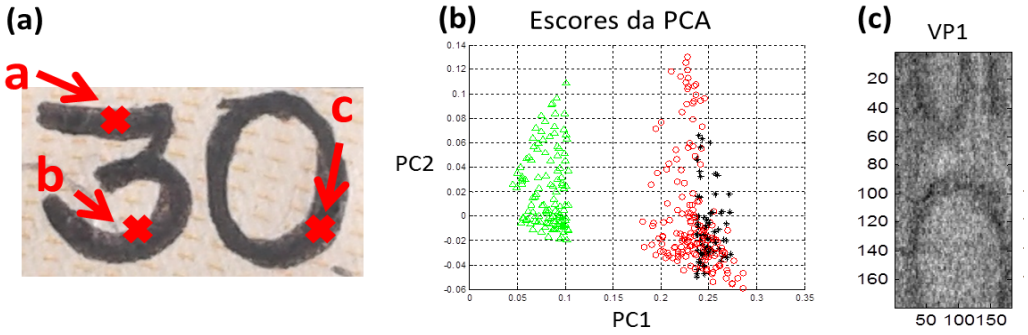


Figura 54: (a) amostra TCF8 com indicações dos pontos amostragem no MIR, (b) gráfico dos escores PCA e (c) imagem dos escores do VP1 de PP.

Amostra TCF9

Para a amostra TCF9 foram realizadas aquisições de imagens na região do MIR em dois pontos do número **60** registrado em papel de cheque. A primeira avaliação dos dados mostrou que menos de 60 pixels de cada imagem foram preservados após a ROI, por esse motivo as aquisições foram refeitas em triplicata para compor os dois conjuntos de dados de cada algarismo do número **60** da amostra. Mesmo após usar um conjunto de dados maior, restaram poucos pixels de cada conjunto de dados. Após aplicar a PCA, os dois conjuntos foram totalmente dissociados e esse resultado foi confirmado pela análise de PP. Os resultados obtidos usando a PCA e PP pra a análise da imagem no NIR se opõem aos anteriores, uma vez que o número **60** completo foi detectado nas imagens dos escores da PC2, da PC5, do VP2 e do VP5. De acordo com o colaborador F a amostra TCF9 contém o número **60** escrito com uma única caneta, portanto o método de resolução do caso usando PCA e HSI-MIR representa um falso-positivo.

Amostra TCF10

A amostra TCF10 foi submetida às técnicas de aquisição de imagem, no infravermelho médio e próximo, associadas à PCA e PP. Entretanto não houve discriminação entre nenhum dos conjuntos de dados no MIR. Além disso, as imagens dos escores da PCA e de PP não mostraram diferenças entre a tinta e o papel. Dessa forma pode-se dizer que as técnicas de reconhecimento de padrão falharam em detectar a falsificação feita pelo colaborador F.

A Tabela 2 apresenta um resumo dos resultados obtidos nos testes-cego. Alguns teste-cego foram solucionados apenas com uma das técnicas de aquisição de imagens utilizadas e apenas 3 testes não foram solucionados com ambas as técnicas de imagem e de análise multivariada utilizadas.

Foram preparados 6 testes-cego pelo colaborador C dos quais apenas o TCC3, que era resultado da combinação das canetas G1 e G3, não foi solucionado. Na etapa de discriminação de tintas de caneta, descrita no capítulo 3, essa combinação foi discriminada com PCA.

Os testes-cego preparados pelo colaborador A usando à combinação da caneta E4 com as canetas E2, E3, E5 e E9 não foram solucionados com as imagens no MIR. Em todos os testes da análise discriminante a caneta E4 apresentou dificuldade de separação e apenas PP foi capaz de diferencia-la das outras canetas esferográficas. A combinação de E4 e E9 no teste-cego TCA4 não teve solução mesmo quando foi usada HSI-NIR. O teste-cego TCA6 também não foi solucionado usando HSI-MIR, essa amostra é resultado da combinação da tinta das canetas E1 e E2 que foi discriminada apenas pela análise de PP dos dados no MIR.

O teste-cego TCF1 também foi preparado através da combinação das canetas E4 e E9, no entanto usando HSI-MIR foi possível determinar a falsificação com PCA. Usando HSI-NIR não foi possível diferenciar o papel da tinta. O teste-cego TCF10, que foi preparada pela combinação das canetas G3 e R2, não foi solucionado por nenhum dos métodos aplicados. A dificuldade de diferenciar a tinta das canetas G3 e R2 já havia sido observada na etapa discriminante, em que apenas PP foi capaz de discriminar os dois conjuntos de dados usando MIR.

Tabela 2: Resumo das soluções obtidas para os testes-cego. Onde: “Sim” significa que a técnica obteve sucesso; “Não” significa que a técnica não obteve sucesso; “Indef.” significa que os resultados obtidos não contribuíram para chegar à solução.

Código da Amostra	**HSI-MIR (met. 1)**		**HSI-NIR (met. 2)**		**Observações e Conclusão**
	PCA	**PP**	**PCA**	**PP**	
TCC1	Indef.	Indef.	Sim	Sim	Teste solucionado
TCC2	Sim	Sim	Sim	Sim	Teste solucionado
TCC3	Não	Não	Indf.	Indf.	Met1- não detectou falsificação/ met.2 não diferenciou tinta e o papel.
TCC4	Sim	Sim	Sim	Sim	Teste solucionado
TCC5	Sim	Sim	Sim	Sim	Teste solucionado
TCC6	Sim	Sim	Sim	Sim	Teste solucionado
TCA1	Indef.	Indef.	Sim	Sim	Met.1 usou poucos pixels/ Teste solucionado

TCA2	Não	Não	Sim	Sim	Met.1 não detectou falsificação/Teste solucionado
TCA3	Sim	Sim	Indef.	Indef.	Met.2 não distinguiu a tinta e o papel/Teste solucionado
TCA4	Não	Não	Não	Não	Nenhum dos métodos detectou a falsificação
TCA5	Sim	Sim	Sim	Sim	Teste solucionado
TCA6	Não	Não	Sim	Sim	Met.1 não detectou falsificação/Teste solucionado
TCA7	Não	Sim	Sim	Sim	Apenas PP do met.1 detectou falsificação/Teste solucionado
TCA8	Sim	Sim	Sim	Sim	Teste solucionado
TCA9	Sim	Sim	Sim	Sim	Teste solucionado
TCA10	Sim	Sim	Sim	Sim	Teste solucionado
TCA11	Sim	Sim	Sim	Sim	Teste solucionado
TCA12	Indef.	Indef.	Sim	Sim	Met.1 usou poucos pixels/Teste solucionado
TCA13	Indef.	Indef.	Sim	Sim	Met.1 usou poucos pixels/Teste solucionado
TCA14	Não	Sim	Não	Sim	Apenas PP detectou adulteração/Teste solucionado
TCF1	Sim	Sim	Indef.	Indef.	Met.2 não distinguiu a tinta e o papel/Teste solucionado
TCF2	Não	Sim	Sim	Sim	Apenas PP do met.1 detectou falsificação/Teste solucionado
TCF3	Sim	Sim	Sim	Sim	Teste solucionado
TCF4	Sim	Sim	Sim	Sim	Teste solucionado
TCF5	Não	Sim	Sim	Sim	Apenas PP do met.1

					detectou falsificação/Teste solucionado
TCF6	Sim	Sim	Sim	Sim	Teste solucionado
TCF7	Sim	Sim	Sim	Sim	Teste solucionado
TCF8	Não	Não	Não	Sim	Met.1 caracterizou falso-positivo/Teste solucionado
TCF9	Não	Não	Sim	Sim	Met.1 caracterizou falso-positivo/Teste solucionado
TCF10	Não	Não	Não	Não	Nenhum dos dois métodos identificou a falsificação

Para compreender melhor como é feita a identificação de falsificação nos casos de teste-cego, é necessário avaliar a influência das variáveis espectrais na discriminação entre a tinta e o papel e entre uma tinta e outra.

A verificação de fraudes usando HSI-MIR é feita por análise das diferenças encontradas entre os pixels mais relacionados com a tinta, portanto é possível verificar qual região do espectro é mais importante para discriminação de grupos usando a matriz de pesos da PCA. A Figura 55a e b mostra o gráfico de escores e de pesos para a PCA usada na resolução do teste-cego TCA10. É possível observar que a maior influência para discriminação das tintas é dada pelas variáveis contidas entre 900 e 1700 cm^{-1} (região de impressão digital). A banda em 3500 cm^{-1} tem menos influência por estar mais correlacionada com estiramentos $O-H$ da celulose.

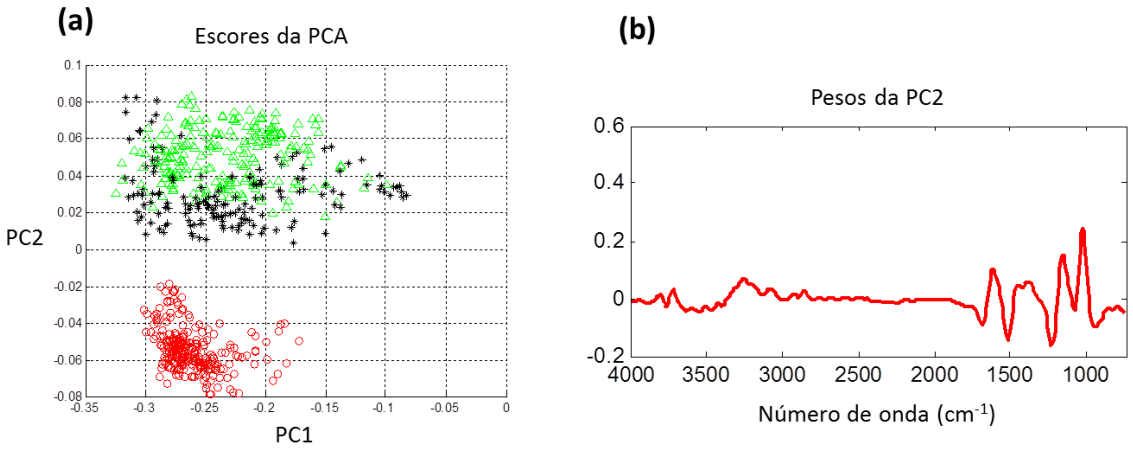


Figura 55: (a) Gráfico dos escores e (b) gráfico dos pesos da PCA para a amostra TCA10.

As imagens hiperespectrais obtidas na região do NIR apresentam regiões com variáveis mais importantes para discriminação da tinta e do papel, bem como variáveis responsáveis pela discriminação das tintas de caneta diferentes usadas para produzir as amostras do teste-cego. Como exemplo, a Figura56a-b mostra o comportamento do gráfico dos pesos na identificação da adulteração observada na constituição da amostra TCA8. Como pode ser visto na Figura 56a, a PC1 mostra um contraste entre as duas tintas. Na Figura 56b, observa-se que as variáveis com loadings mais negativos discriminam a tinta usada para realizar os traços horizontais que correspondem a adulteração (variáveis iniciais e após 2300 nm) e os as variáveis com loadings mais positivos estão mais relacionadas com a tinta do traço vertical.

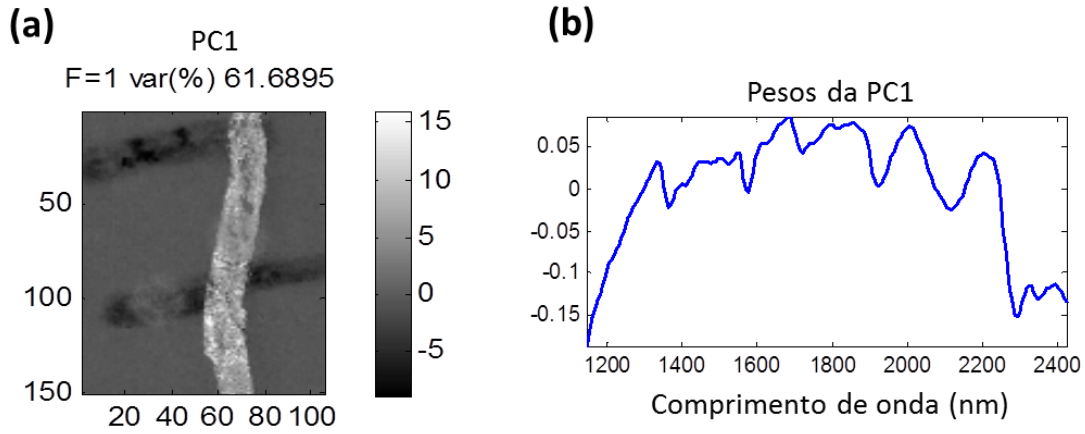


Figura 56: (a) Imagem dos escores da PC1, (b) gráfico dos pesos da PC1.

8 CONCLUSÃO TESTE - CEGO

O melhor pré-processamento para os espectros de infravermelho próximo das amostras de teste-cego foi obtido a partir das correções de efeito aditivo provocado pelo espalhamento da radiação. Após avaliar algumas técnicas de otimização dos espectros, SNV foi definida como a mais eficiente na correção dos efeitos indesejados e com melhores resultados para a separação entre o papel e as tintas.

Em três amostras um grande número de pixels, adquiridos com HSI-MIR, foram associados ao papel. Isso implica dizer que poucos pixels da imagem serão utilizados para diferenciar as duas possíveis tintas presentes na amostra comprometendo a acurácia do método. Por esse motivo, para essas três amostras o resultado foi considerado indefinido por não haver informação suficiente para inferir sobre a composição da amostra. As imagens no NIR também apresentaram problemas na separação do papel e das tintas para três amostras. Nesses casos não foi possível inferir sobre a disposição das tintas, portanto os resultados foram considerados indefinidos.

Com base nos resultados obtidos, a combinação de duas canetas de tinta a base de óleo (canetas esferográficas) representa a maior dificuldade na detecção das fraudes usando HSI-MIR. Cinco amostras dos testes-cego preparados pelo colaborador A que continham adulteração não foram solucionadas com HSI-MIR, todas as combinações de canetas esferográficas. Tendo em vista os resultados apresentados, ficou claro que a melhor forma de avaliar amostras é associar as duas técnicas de aquisição de imagens hiperespectrais apresentadas com as técnicas de reconhecimento de padrão PCA e PP. Os resultados foram melhores quando a análise de PP foi implementada na avaliação das imagens nas duas regiões espectrais. Enquanto a PCA obteve 50% e 76,7 % de acerto na resolução dos testes-cego para nas regiões do MIR e NIR, respectivamente, a análise de PP obteve correta determinação em 63,3% dos testes no MIR e em 83,3% dos testes no NIR.

O maior percentual de correta determinação foi de 83,3% usando as imagens hiperespectrais no NIR, no entanto, entre os casos não resolvidos com esse método, dois testes-cego foram solucionados usando HSI-MIR. Dessa forma, a aplicação análise de PP nas imagens hiperespectrais no infravermelho médio e próximo resultou na correta resolução de 90% dos testes-cego.

9 CONSIDERAÇÕES FINAIS

A primeira etapa deste trabalho foi importante para estabelecer as metodologias de análise dos dados de imagens hiperespectrais no MIR. Esse primeiro desafio foi investigar a dissimilaridades na composição química de diferentes tintas de caneta preta, uma vez que estas tintas são muito semelhantes ou até idênticas diante de inspeção visual. Para demonstrar a capacidade de aquisição de informação química da técnica de imagem por ATR e o poder discriminante das técnicas de reconhecimento de padrão PCA e PP, foram produzidas120 combinações de 16 canetas de tinta preta, em que 89 pares de imagens foram discriminados usando PCA e 117 pares usando a análise de PP nos conjuntos de dados. Esses testes mostraram que as técnicas de aquisição e discriminação têm potencial para serem aplicadas em testes reais e essa conclusão reforçou os objetivos da segunda etapa.

A elaboração de uma metodologia para abordagem de amostras de texto desconhecidas é uma necessidade na investigação forense de documentos. Nesse sentido foram avaliadas 30 amostras preparadas sigilosamente por colaboradores, simulando casos reais, das quais 15 foram corretamente descritas usando PCA e HSI-MIR e 23 usando PCA e HSI-NIR. Empregando análise de PP os resultados foram superiores, com 19 amostras descritas corretamente usando PP e HSI-MIR e 25 amostras usando PP e HSI-NIR. Um fator que influenciou o desempenho inferior no uso de imagens hiperespectrais no MIR foi o número de amostragens feitas usando a técnica de aquisição de imagens por ATR. A aquisição das imagens HSI-MIR também é mais demorada e a necessidade de um contato perfeito entre a amostra e o cristal faz com que aconteça certa deformação do mesmo. Mas a técnica também se provou ser uma poderosa ferramenta de análise de tinta em papel.

O conjunto de todos os resultados mostra que HSI-NIR e HSI-MIR associadas a técnicas de reconhecimento de padrão têm um grande potencial de aplicação em rotinas de análise da autenticidade de documentos.

10 PERSPECTIVAS FUTURAS

Estudos futuros e ainda mais completos sobre a temática abordada neste trabalho são necessários para aprimorar a metodologia de análise e ampliar as fronteiras de aplicação. É possível otimizar a metodologia para torná-la menos subjetiva e também mais robusta aumentando o número de amostragens nas imagens obtidas na região MIR. Pode-se vislumbrar o aumento do conjunto de canetas e adicionar maior variabilidade de marcas e modelos em cada amostras. Também pode ser investigada a influência da textura do papel e registros impressos, tais como marca d’agua em talão de cheque e em outros documentos oficiais.

REFERÊNCIAS

ADAM, C. D.; SHERRATT, S. L.; ZHOLOBENKO, V. L. Classification and individualization of black ballpoint pen inks using principal component analysis of UV–vis absorption spectra. **Forensic Science International**. v. 174, p. 16–25, 2008.

AMIGO, J. M.; RAVN, C. Direct quantification and distribution assessment of major and minor components in pharmaceutical tablets by NIR-chemical imaging. **European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences**, v. 37, n. 2, p. 76–82, 2009.

BEEBE, K. R., PELL, R. J., SEASHOLTZ, M. B. **Chemometrics: a Practical Guide**. Nova Iorque: Wiley-Interscience. 1998, 360 p.

BOJKO, K.; ROUX, C.; REEDY, B. J. An examination of the sequence of intersecting lines using attenuated total reflectance-Fourier transform infrared spectral imaging. **Journal of forensic sciences**, v. 53, n. 6, p. 1458–1467, 2008.

BRAZ, A.; LÓPEZ-LÓPEZ, M.; GARCÍA-RUIZ, C. Raman spectroscopy for forensic analysis questioned documents. **Forensic Science International**, v. 232, p. 206−212, 2013.

BRAZ, A. Investigating current challenges in forensic ink analysis by raman spectroscopy. 2014. 199 f. Tese (Doutorado em criminalistica) – Instituto Universitário de Investigación en Ciencias Policiales, Universidad de Alcalá, Alcalá. 2014.

BRAZ, A.; LÓPEZ-LÓPEZ, M.; GARCÍA-RUIZ, C. Raman imaging for determining the sequence of blue pen ink crossings. **Forensic Science International**, v. 249, p. 92−100, 2015.

BRUNELLE, R.L.; CRAWFORD, K.R. **Advances in the Forensic Analysis and Dating of Writing Ink**. Springfield. Charles C Thomas Pub Ltd, 2003. 236 p.

CHEN, T.; SCHULTZ, Z. D.; LEVIN, I. W. Infrared spectroscopy imaging of latent fingerprints and associated forensic evidences. **Analyst,** v. 139, p. 1902-1904, 2009.

CRUZ, J.; BAUTISTA, M.; AMIGO, J. M.; BLANCO, M. Nir-chemical imaging study of acetylsalicylic acid in commercial tablets. **Talanta,** v. 80, n. 2, p. 473–478, 2009.

DE JUAN, A; MAEDER, M; HANCEWICZ, T; DUPONCHEL, L; TAULER, R. Chemometric Tools for Image Analysis. In: SALZER, R. e SIESLER, H. W. **Infrared and Raman Spectroscopic Imaging** 1. ed. Alemanha: Wiley-VCH, 2009. p. 65-109.

ELLISON, C. D.; ENNIS, B. J.; HAMAD, M. L.; LYON, R. C. Measuring the distribution of density and tabletting force in pharmaceutical tablets by chemical imaging. **Journal of pharmaceutical and biomedical analysis,** v. 48, n. 1, p. 1–7, 2008.

EZCURRA, M.; GÓNGORA, J. M. G.; MAGUREGUI, I.; ALONSO, R. Analytical methods for dating modern writing instrument inks on paper. **Forensic Science International**, v. 197, p. 1–20, 2010.

FEARN, T.; RICCIOLI C.; GARRIDO-VARO, A.; GUERRERO-GINEL, J. E. On the geometry of SNV and MSC. **Chemometrics and Intelligent Laboratory Systems**, v. 96, n. 1, p. 22-26, 2009.

FRIEDMAN, J. H., TUKEY, J.W. A projection pursuit algorithm for exploratory data analysis. **IEE Transactions on Computers**, v. c-23, n. 9, p. 881-890, 1974.

GONZALES, R. C.; WOODS, R. E. **Digital image processing.** 2. ed. New Jersey: Prentice Hall, 2001. p. 15-45.

GOWEN, A. A.; DONNELL, C. P. O.; CULLEN, P. J.; DOWNEY, G.; FRIAS, J. M. Hyperspectral imaging e an emerging process analytical tool for food quality and safety control. **Trends in Food Science & Technology,** v. 18, n. 12, p. 590-598, 2007.

HOU, S.; WENZTELL, P. D. Regularized projection pursuit for data with a small sample-to-variable ratio. **Metabolomics**. v. 10, n. 4, p. 589-606, 2013.

HOU, S. WENTZELL, P. D. Fast and simple methods for the optimization of kurtosis used as a projection pursuit index. **Analytica Chimica Acta**.v. 704, p. 1– 15, 2011.

IGOE, T. J.; REYNOLDS, B. L. A Lifting Process For Determining The Writing Sequence Of Two Intersecting Ball-Point Pen Strokes. **Forensic Science International**, v. 20, p. 201–205, 1982.

LI, B.; BEVERIDGE, P.; O'HARE, W. T.; ISLAM, M. The age estimation of blood stains up to 30 days old using visible wavelength hyperspectral image analysis and linear discriminant analysis. **Science and Justice**. v. 53, p. 270–277, 2013.

LING, C.; SOMMER, A. The Advantages of an Attenuated Total Internal Reflection Infrared Microspectroscopic Imaging Technique for the Analysis of Polymer Laminates. **Microscopy and Microanalysis,** v. 21, p. 626-636, 2015.

LUO, J.; YING, K.; BAI, J. Savitzky-Golay smoothing and differentiation filter for even number data. **Signal Processing**, v. 85, p. 1429-1434, 2005.

MALINEN, J.; SAARI, H.; KEMENY, G.; SHI, G.; ANDERSON, C. Comparative performance studies between tunable filter and pushbroom chemical imaging systems. **Proc. SPIE 7680, Next-Generation Spectroscopic Tecnologies III**, v. 7680 ed, 2010; doi**;** doi:10.1117/12.850105; Disponível em: < http://proceedings.spiedigitallibrary.org/volume.aspx?volumeid=830>. Acesso 08 de ago. 2015.

MURO, C. K.; DOTY, K. C.; BUENO, J.; HALÁMKOVÁ, L.; LEDNEV, I. K. Vibrational Spectroscopy: Recent Developments to Revolutionize Forensic Science. **Analytical Chemistry**, v. 87, p. 306-327, 2015

NAM, Y. S.; PARK, J. S.; LEE, Y.; LEE, K. B. J. Application of Micro-Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy to Ink Examination in Signatures Written With Ballpoint Pen on Questioned Documents. **Forensic Sciences**. V. 59, p 800−805, 2014.

PASQUINI, C. Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications. **Journal of Brazilian Chemical Society**, v. 14, n. 2, p. 198-219, 2003.

PAVIA, D. L.; LAMPMAN, G. M.; KRIZ, G. S.; VYVYAN, J. R. **Introduction to spectroscopy**. 4.ed, Belmont: Brooks/cole, 2009. p. 15-104.

PEÑA, D.; PRIETO, F. J. Cluster Identification Using Projections. **Journal of the American Statistical Association**. v. 96, n. 456, p. 1433-1445, 2001a.

PEÑA, D.; PRIETO, F. J. Multivariate Outlier Detection and Robust Covariance Matrix Estimation. **Technometrics**.v. 43, n. 3, p. 286-310, 2001b.

PRATS-MONTALBÁN, J. M.; DE JUAN, A.; FERRER, A. Multivariate image analysis: A review with applications. **Chemometrics and Intelligent Laboratory Systems**, v. 107, n. 1, p. 1–23, 2011.

RINNAN, Å.; BERG, F. V. D.; ENGELSEN, S. B. Review of the most common pre-processing techniques for near-infrared spectra. **TrAC Trends in Analytical Chemistry**, v. 28, n. 10, p. 1201-1222, 2009.

SILVA, C. S.; PIMENTEL, M. F.; HONORATO, R. S.; PASQUINI, C.;PRATS-MONTALBÁN, J. M.; FERRER, A. Near infrared hyperspectral imaging for forensic analysis of document forgery. **Analyst,** v. 139, p. 5176-5184, 2014a.

SILVA, C. S.; BORBA, F. S. L.; PIMENTEL, M. F.; PONTES, M. J. C.; HONORATO, R. S.; PASQUINI, C. Classification of blue pen ink using infrared spectroscopy and linear discriminant analysis. **Microchemical Journal,** v. 109, p. 122–127, 2013.

SILVA, V. A. G.; TALHAVINI, M.; ZACCA, J. J.; TRINDADE, B. R.; BRAGA, J. W. B. Discrimination of Black Pen Inks on Writing Documents Using Visible Reflectance Spectroscopy and PLS-DA. **Journal of the Brazilian Chemical Society,** v. 25, n. 9, p. 1552-1564, 2014b.

SILVA, V. A. G.; TALHAVINI, M.; ZACCA, J. J.; TRINDADE, B. R.; BRAGA, J. W. B. Non-destructive identification of different types and brands of blue pen inks in cursive handwriting by visible spectroscopy and PLS-DA for forensic analysis. **Microchemical Journal,** v. 116, p. 235-243, 2014c.

SILVERSTEIN, R. M.; WEBSTER, F. X.; KIEMLE, D. J. **Spectrometric Identification of Organic Compounds**. 7. ed. Danvers: John Wiley & Sons, 2005. p. 72-126.

SKOOG, D. A.; HOLLER, F. J.; CROUCH, S. R. **Princípios de Análise Instrumental**.6. ed. Porto Alegre: Bookman, 2009. p. 444-494.

TAHTOUH, M.; DESPLAND, P.; SHIMMON, R.; KALMAN, J. R.; REEDY, B. J. The application of infrared chemical imaging to the detection and enhancement of latent fingerprints: method optimization and further findings. **Journal of forensic sciences**, v. 52, n. 5, p. 1089–1096, 2007.

THANASOULIAS, N. C.; PARISIS, N. A.; EVMIRIDIS, N. P. Multivariate chemometrics for the forensic discrimination of blue ball-point pen inks based on their Vis spectra. **Forensic Science International**. v. 138, p. 75–84, 2003.

VIDAL, M.; AMIGO, J. M. Pre-processing of hyperspectral images. Essential steps before image analysis. **Chemometrics and Intelligent Laboratory Systems**. v. 117, p. 138–148, 2012.

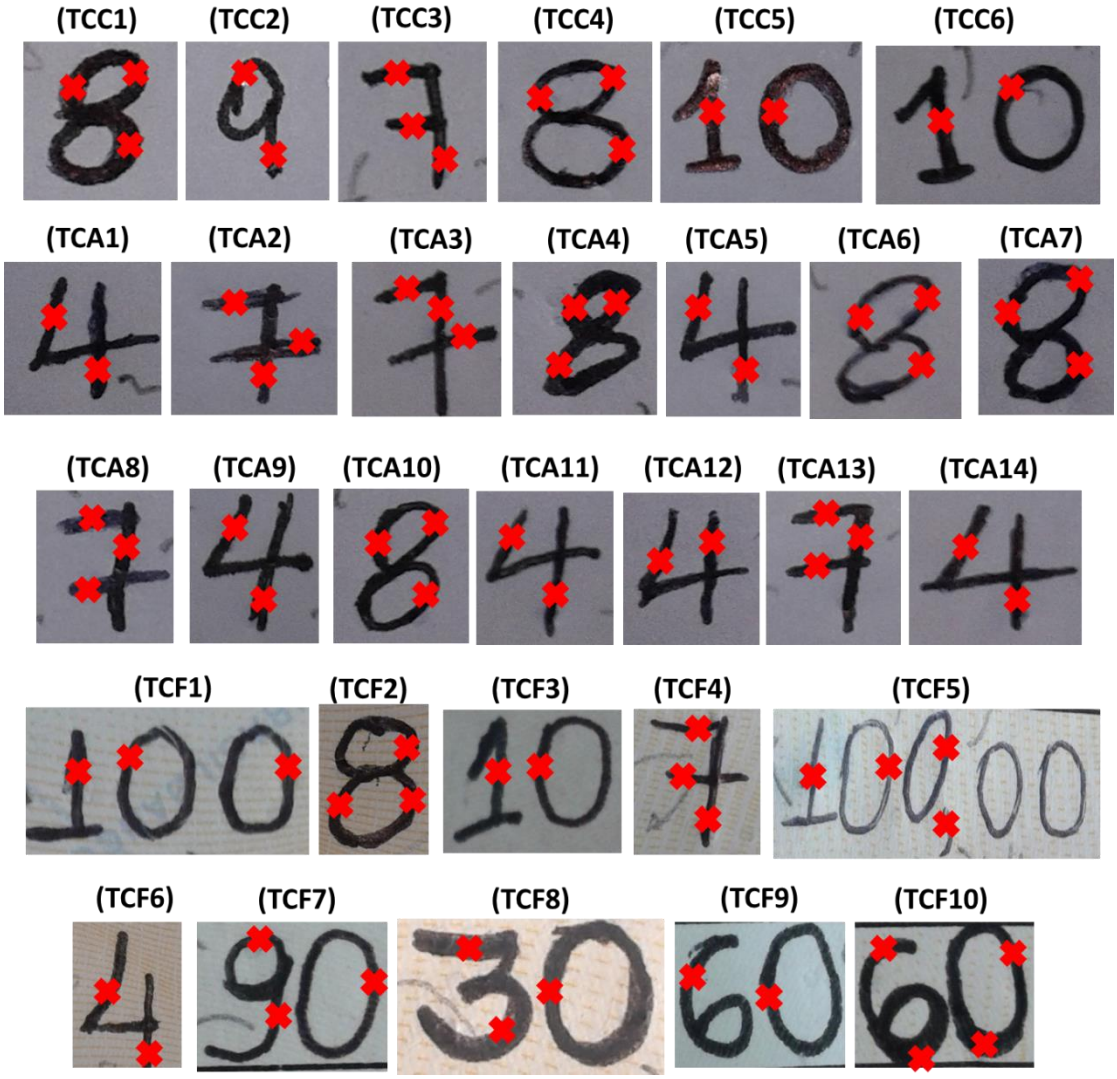
WENTZELL, P. D.; HOU, S.; SILVA, C. S.; WICKS, C. C.; PIMENTEL, M. F. Procrustes rotation as a diagnostic tool for project pursuit analysis. **Analytica Chimica Acta,** v. 877, p. 51-63, 2015.

WENTZELL, P.; HOU, S.; SILVA, C. S.; PIMENTEL, M. F. **Projection Pursuit Revisited: Exploratory Analysis of Multiclass Data**. 2014. Trabalho apresentado no XIV Chemometrics in Analytical Chemistry, Richmond, 2014.

WENTZELL, P. D.; HOU, S. Exploratory data analysis with noisy measurements. **Jornal of Chemometrics**, v. 26, p. 264-281, 2012.

WOLD, S.; ESBENSEN, K.; GELADI, P. Principal Component Análysis. **Chemometrics and Intelligent Laboratory Systems**, v. 2, p. 37-52, 1987.

APÊNDICE A: Foto das amostras de teste-cego preparadas pelos três colaboradores com as indicações dos pontos de amostragem das imagens hiperespectrais no MIR. As aquisições de imagens hiperespectrais no NIR foram realizadas em toda a área das imagens abaixo.



Fonte: Do autor.

APÊNDICE B: Detalhamento das combinações produzidas com canetas de tintas similares e resultado da análise discriminante. O número de PC's usadas para PP está indicado apenas nos casos em que a PCA não conseguiu discriminar as duas canetas.

Tipo de canetas Combinadas	Combinação		Resultados	
	Caneta 1	Caneta 2	PCA	PP
Óleo x Óleo	E1	E2	Não	Sim (14PC's)
Óleo x Óleo	E1	E3	Sim	Sim
Óleo x Óleo	E1	E4	Sim	Sim
Óleo x Óleo	E1	E5	Sim	Sim
Óleo x Óleo	E1	E6	Não	Sim (16 PC's)
Óleo x Óleo	E1	E7	Sim	Sim
Óleo x Óleo	E1	E8	Não	Sim (16 PC's)
Óleo x Óleo	E1	E9	Não	Sim (4 PC's)
Óleo x Óleo	E2	E3	Sim	Sim
Óleo x Óleo	E2	E4	Não	Sim (6 PC's)
Óleo x Óleo	E2	E5	Sim	Sim
Óleo x Óleo	E2	E6	Não	Não
Óleo x Óleo	E2	E7	Sim	Sim
Óleo x Óleo	E2	E8	Não	Sim (7 PC's)
Óleo x Óleo	E2	E9	Não	Sim (6 PC's)
Óleo x Óleo	E3	E4	Não	Sim (19 PC's)
Óleo x Óleo	E3	E5	Sim	Sim
Óleo x Óleo	E3	E6	Sim	Sim
Óleo x Óleo	E3	E7	Sim	Sim
Óleo x Óleo	E3	E8	Sim	Sim
Óleo x Óleo	E3	E9	Sim	Sim
Óleo x Óleo	E4	E5	Não	Sim (6 PC's)

Óleo x Óleo	E4	E6	Não	Sim (6 PC's)
Óleo x Óleo	E4	E7	Sim	Sim
Óleo x Óleo	E4	E8	Sim	Sim
Óleo x Óleo	E4	E9	Não	Sim (13 PC's)
Óleo x Óleo	E5	E6	Sim	Sim
Óleo x Óleo	E5	E7	Sim	Sim
Óleo x Óleo	E5	E8	Sim	Sim
Óleo x Óleo	E5	E9	Não	Sim (6 PC's)
Óleo x Óleo	E6	E7	Sim	Sim
Óleo x Óleo	E6	E8	Não	Sim (13 PC's)
Óleo x Óleo	E6	E9	Sim	Sim
Óleo x Óleo	E7	E8	Sim	Sim
Óleo x Óleo	E7	E9	Não	Sim (24 PC's)
Óleo x Óleo	E8	E9	Sim	Sim
Gel x Gel	G1	G2	Sim	Sim
Gel x Gel	G1	G3	Sim	Sim
Gel x Gel	G2	G3	Sim C's)	Sim
Hid. x Hid.	H1	H2	Sim	Sim
Hid. x Hid.	H1	R1	Sim	Sim
Hid. x Hid.	H1	R2	Sim	Sim
Hid. x Hid.	H2	R1	Sim	Sim
Hid. x Hid.	H2	R2	Sim	Sim
Hid. x Hid.	R1	R2	Não	Não (30 PC's)

APÊNDICE C: Detalhamento das combinações produzidas com canetas de tintas diferentes e resultado da analise discriminante. O número de PC's usadas para PP está indicado apenas nos casos em que a PCA não conseguiu discriminar as duas canetas.

Tipos de canetas Combinadas	Combinação		Resultados	
	Caneta 1	Caneta 2	PCA	PP
Esf. x Gel	E1	G1	Não	Sim (13PC's)
Esf. x Gel	E1	G3	Sim	Sim
Esf. x Gel	E1	G2	Sim	Sim
Esf. x Gel	E2	G1	Sim	Sim
Esf. x Gel	E2	G3	Sim	Sim
Esf. x Gel	E2	G2	Sim	Sim
Esf. x Gel	E3	G1	Sim	Sim
Esf. x Gel	E3	G3	Não	Sim (5 PC's)
Esf. x Gel	E3	G2	Sim	Sim
Esf. x Gel	E4	G1	Sim	Sim
Esf. x Gel	E4	G3	Não	Sim (20 PC's)
Esf. x Gel	E4	G2	Sim	Sim
Esf. x Gel	E5	G1	Sim	Sim
Esf. x Gel	E5	G3	Não	Sim (6 PC's)
Esf. x Gel	E5	G2	Sim	Sim
Esf. x Gel	E6	G1	Não	Sim (14 PC's)
Esf. x Gel	E6	G3	Sim	Sim
Esf. x Gel	E6	G2	Sim	Sim
Esf. x Gel	E7	G1	Sim	Sim
Esf. x Gel	E7	G3	Sim	Sim
Esf. x Gel	E7	G2	Sim	Sim
Esf. x Gel	E8	G1	Não	Sim (8 PC's)
Esf. x Gel	E8	G3	Sim	Sim
Esf. x Gel	E8	G2	Sim	Sim

Esf. x Gel	E9	G1	Sim	Sim
Esf. x Gel	E9	G3	Sim	Sim
Esf. x Gel	E9	G2	Sim	Sim
Esf. x Hid.	E1	H1	Sim	Sim
Esf. x Hid.	E1	H2	Sim	Sim
Esf. x Hid.	E1	R1	Sim	Sim
Esf. x Hid.	E1	R2	Sim	Sim
Esf. x Hid.	E2	H1	Sim	Sim
Esf. x Hid.	E2	H2	Sim	Sim
Esf. x Hid.	E2	R1	Não	Sim (8PC's)
Esf. x Hid.	E2	R2	Não	Não (28 PC's)
Esf. x Hid.	E3	H1	Sim	Sim
Esf. x Hid.	E3	H2	Sim	Sim
Esf. x Hid.	E3	R1	Sim	Sim
Esf. x Hid.	E3	R2	Sim	Sim
Esf. x Hid.	E4	H1	Sim	Sim
Esf. x Hid.	E4	H2	Sim	Sim
Esf. x Hid.	E4	R1	Sim	Sim
Esf. x Hid.	E4	R2	Sim	Sim
Esf. x Hid.	E5	H1	Sim	Sim
Esf. x Hid.	E5	H2	Sim	Sim
Esf. x Hid.	E5	R1	Sim	Sim
Esf. x Hid.	E5	R2	Sim	Sim
Esf. x Hid.	E6	H1	Sim	Sim
Esf. x Hid.	E6	H2	Sim	Sim
Esf. x Hid.	E6	R1	Sim	Sim
Esf. x Hid.	E6	R2	Não	Sim (8PC's)
Esf. x Hid.	E7	H1	Sim	Sim
Esf. x Hid.	E7	H2	Sim	Sim
Esf. x Hid.	E7	R1	Sim	Sim

Esf. x Hid.	E7	R2	Sim	Sim
Esf. x Hid.	E8	H1	Sim	Sim
Esf. x Hid.	E8	H2	Sim	Sim
Esf. x Hid.	E8	R1	Sim	Sim
Esf. x Hid.	E8	R2	Não	Sim (8PC's)
Esf. x Hid.	E9	H1	Sim	Sim
Esf. x Hid.	E9	H2	Sim	Sim
Esf. x Hid.	E9	R1	Sim	Sim
Esf. x Hid.	E9	R2	Não	Sim (4 PC's)
Gel x Hid.	G1	H1	Sim	Sim
Gel x Hid.	G1	H2	Sim	Sim
Gel x Hid.	G1	R1	Não	Sim (4PC's)
Gel x Hid.	G1	R2	Sim	Sim
Gel x Hid.	G3	H1	Sim	Sim
Gel x Hid.	G3	H2	Sim	Sim
Gel x Hid.	G3	R1	Não	Sim (7 PC's)
Gel x Hid.	G3	R2	Não	Sim (5 PC's)
Gel x Hid.	G2	H1	Sim	Sim
Gel x Hid.	G2	H2	Sim	Sim
Gel x Hid.	G2	R1	Sim	Sim
Gel x Hid.	G2	R2	Não	Sim (12 PC's)