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

TESE DE DOUTORADO

SELEÇÃO DE VARIÁVEIS NO DESENVOLVIMENTO, CLASSIFICAÇÃO E PREDIÇÃO DE PRODUTOS

Karina Rossini

Porto Alegre, 2011

Karina Rossini

SELEÇÃO DE VARIÁVEIS NO DESENVOLVIMENTO, CLASSIFICAÇÃO E PREDIÇÃO DE PRODUTOS

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

Orientador: Prof. Flávio Sanson Fogliatto, Phd.

Prof. El Mostafa Qannari, Phd.

Porto Alegre, 2011

Karina Rossini

SELEÇÃO DE VARIÁVEIS NO DESENVOLVIMENTO, CLASSIFICAÇÃO E PREDIÇÃO DE PRODUTOS

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

_____

Prof. Flávio Sanson Fogliatto, Phd.

PPGEP / UFRGS

Orientador

_____

Prof. ª Carla Schwengberg ten Caten, Dra.

Coordenador PPGEP / UFRGS

Banca Examinadora:

Profa. Linda Lee Ho, Dra. (Dep. de Eng. de Produção / USP)

Profa. Simone Hickmann Flôres, Dra. (Dep.de Eng. de Alimentos / UFRGS)

Prof. Michel Anzanello, Phd. (Dep. de Eng. de Produção / UFRGS)

Dedicatória

Ao meu marido Juliano e

aos meus amados pais Alzira e Arcelo.

AGRADECIMENTOS

Ao meu marido Juliano e aos meus pais Alzira e Arcelo, pela compreensão, incentivo, carinho e amor incondicional.

Aos meus irmãos Gustavo e Cassiane, sobrinhos Vinícius e Fernando, e familiares pelo carinho e estímulo.

Aos meus orientadores, o professor Flogliatto e professor Qannari, pela ajuda e orientação constante.

Aos colegas do LOPP pelo companheirismo e amizade.

Ao bolsista Felipe pela dedicação e interesse neste trabalho.

Aos colegas da École National Vétérinarie Agroalimentaire et de l'Alimentation pelo apoio em momentos difíceis.

Aos membros da banca examinadora pelas contribuições e críticas ao trabalho.

A todos que de alguma forma contribuíram para o desenvolvimento e realização deste trabalho.

ROSSINI, Karina ***Seleção de variáveis no desenvolvimento, classificação e predição de produtos*****,** 2011. Tese (Doutorado em Engenharia) – Universidade Federal do Rio Grande do Sul, Brasil.

RESUMO

O presente trabalho apresenta proposições para seleção de variáveis em avaliações sensoriais descritivas e de espectro infravermelho que contribuam com a indústria de alimentos e química através da utilização de métodos de análise multivariada. Desta forma, os objetivos desta tese são: (*i*) Estudar as principais técnicas de análise multivariada de dados, como são comumente organizadas e como podem contribuir no processo de seleção de variáveis; (*ii*) Identificar e estruturar técnicas de análise multivariada de dados de forma a construir um método que reduza o número de variáveis necessárias para fins de caracterização, classificação e predição dos produtos; (*iii*) Reduzir a lista de variáveis/atributos, selecionando aqueles relevantes e não redundantes, reduzindo o tempo de execução e a fadiga imposta aos membros de um painel em avaliações sensoriais; (*iv*) Validar o método proposto utilizando dados reais; e (*v*) Comparar diferentes abordagens de análise sensorial voltadas ao desenvolvimento de novos produtos. Os métodos desenvolvidos foram avaliados através da aplicação de estudos de caso, em exemplos com dados reais. Os métodos sugeridos variam com as características dos dados analisados, dados altamente multicolineares ou não e, com e sem variável dependente (variável de resposta). Os métodos apresentam bom desempenho, conduzindo a uma redução significativa no número de variáveis e apresentando índices de adequação de ajuste dos modelos ou acurácia satisfatórios quando comparados aos obtidos mediante retenção da totalidade das variáveis ou comparados a outros métodos dispostos na literatura. Conclui-se que os métodos propostos são adequados para a seleção de variáveis sensoriais e de espectro infravermelho.

Palavras-chave: seleção de variáveis, análise sensorial, análise multivariada de dados, espectro infravermelho.

ROSSINI, Karina ***Selection of variables for the development, classification, and prediction of products***, 2011. Dissertation (Doctorate in Engineering) – Universidade Federal do Rio Grande do Sul, Brazil.

ABSTRACT

This dissertation presents propositions for variable selection in data from descriptive sensory evaluations and near-infrared (NIR) spectrum analyses, based on multivariate analysis methods. There are five objectives here: (*i*) review the main multivariate analysis techniques, their relationships and potential use in variable selection procedures; (*ii*) propose a variable selection method based on the techniques in (*i*) that allows product prediction, classification, and description; (*iii*) reduce the list of variables/attributes to be analyzed in sensory panels identifying those relevant and non-redundant, such that the time to collect panel data and the fatigue imposed on panelists is minimized; (*iv*) validate methodological propositions using real life data; and (*v*) compare different sensory analysis approaches used in new product development. Proposed methods were evaluated through case studies, and vary according to characteristics in the datasets analyzed (data with different degrees of multicollinearity, presenting or not dependent variables). All methods presented good performance leading to significant reduction in the number of variables in the datasets, and leading to models with better adequacy of fit. We conclude that the methods are suitable for datasets from descriptive sensory evaluations and NIR analyses.

Keywords: Variable selection, sensory evaluation, multivariate data analysis, near-infrared (NIR) spectrum analyses.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Em ambientes industriais, é comum que processos demandem um grande número de variáveis para seu controle. O processamento eficiente de grandes bases de dados tem representado um desafio crescente para empresas, devido à difusão de sensores e dispositivos automatizados de medição na caracterização de seus processos (KETTANEH et al., 2005). Assim, a identificação das variáveis importantes na descrição de processos tornou-se essencial para o seu controle e otimização eficientes, bem como para a classificação dos produtos que deles emergem.

No âmbito da indústria de alimentos e química, a relevância deste tema concerne, sobretudo, a área de avaliações sensoriais, onde a existência de um elevado número de variáveis reflete em procedimentos longos e caros, que impõem fadiga aos membros do painel, e em análises químicas onde o número de variáveis coletadas é excessivamente grande, como no caso de análises em espectros de infravermelho.

A análise sensorial compreende um conjunto de técnicas para medir precisamente atributos sensoriais de produtos a partir de respostas humanas. Tais técnicas utilizam princípios oriundos da ciência de alimentos, fisiologia, psicologia e estatística, fornecendo respostas objetivas para as propriedades de alimentos, conforme percebidas pelos cinco sentidos (PIGGOTT *et al.*, 1998).

Informações obtidas através de avaliações sensoriais podem ser utilizadas por empresas como suporte técnico para pesquisa, industrialização, marketing e controle de qualidade dos itens produzidos, qualificando decisões técnicas e administrativas. Na perspectiva do consumidor, a avaliação sensorial em produtos industriais assegura que os mesmos cheguem ao mercado com as características desejadas, através da identificação de suas expectativas (LEDAUPHIN *et al.*, 2008; DUTCOSKI, 1996; LAWLESS; HEYMANN, 1998).

A avaliação sensorial é realizada de acordo com diferentes testes que dependem da sua finalidade (ANZALDÚA-MORALES, 1994). Limitações comuns nesses testes referem-se ao número de amostras, número de julgadores ou quantidade de atributos a serem analisados. No que tange a esse último aspecto, Sahmer e Qannari (2008) apontam que a seleção de atributos tem sido foco de investigações em diversas áreas industriais.

Ainda que amplamente utilizados (MURRAY *et al.*, 2001), os métodos de Análise Descritiva (AD) apresentam algumas limitações. Dentre elas, o número de atributos a serem avaliados pelos julgadores costuma ser extenso; alguns protocolos de AD podem apresentar até 30 atributos (CARBONELL *et al.*, 2007).

A seleção de atributos contribui com essa limitação e pode destinar-se tanto a identificar um subconjunto de atributos não redundantes que melhor descrevem os produtos, quanto a encontrar atributos que melhor discriminam os produtos. Um atributo relevante em uma investigação sensorial é aquele cuja avaliação apresenta diferenças sistemáticas e significativas entre as amostras investigadas, tal que seja possível relacionar o nível do atributo (através de seu valor medido) com características das amostras (por exemplo, a presença ou ausência de um ingrediente). Atributos que apresentam tal comportamento poderiam, por exemplo, ser utilizados como variáveis de resposta em modelos de regressão, os quais permitiriam a otimização do produto investigado no painel sensorial. Assim, independente da finalidade da seleção (predição ou classificação), o objetivo é reduzir o número de variáveis necessárias através da eliminação de variáveis de ruído, de variáveis com forte correlação com as demais ou que sejam irrelevantes na caracterização dos produtos.

Em aplicações estatísticas e matemáticas onde o objetivo é encontrar bons modelos para predição de variáveis de resposta, o excessivo número de variáveis preditoras (variáveis independentes) pode resultar na inclusão de variáveis de ruído. Este problema pode ser estendido para a quimiometria e aplicações industriais, onde centenas e até milhares de variáveis são coletadas.

As melhores técnicas estatísticas para selecionar variáveis dependem da aplicação pretendida e da natureza do banco de dados. Uma variedade de métodos tem sido proposta até então com base em paradigmas diversos, tais como procedimentos sequenciais (ANZANELLO *et al.*, 2011), Baumann *et al.* (2002), Roy e Roy (2008), análise de *clusters* (SAHMER E QANNARI, 2008), e algoritmo genético (FERRAND *et al.*, 2010 e LEARDI *et al.*, 2002).

1.1 TEMA E OBJETIVOS

O tema de pesquisa desta tese contempla seleção de variáveis nas áreas de análise sensorial de alimentos e análises de espectro infravermelho, fundamentada em técnicas de análise multivariada e de mineração de dados. Dentro destas áreas, o tema desta pesquisa concentra-se, majoritariamente, na seleção de variáveis que (*i*) identifiquem os atributos com habilidades de discriminar produtos, classificando-os em classes distintas, além de refletir a consistência dos julgadores; e (*ii*) identifiquem variáveis a serem usadas em modelos de predição das caraterísticas dos produtos, particularmente em bancos de dados altamente multicolineares, como no caso daqueles oriundos da análise de espectros de infravermelho.

O objetivo principal deste trabalho é desenvolver metodologias de seleção de variáveis em avaliações sensoriais descritivas e de espectro infravermelho que contribuam com a indústria de alimentos e química através da utilização de métodos de análise multivariada. Como decorrência do objetivo principal, os objetivos específicos descritos a seguir podem ser atingidos:

a) Estudar as principais técnicas de análise multivariada de dados, como são comumente organizadas e como podem contribuir no processo de seleção de variáveis;

b) Identificar e estruturar técnicas de análise multivariada de dados de forma a construir um método que reduza o número de variáveis necessárias para fins de caracterização, classificação e predição dos produtos;

c) Reduzir a lista de variáveis/atributos, selecionando aqueles relevantes e não redundantes, reduzindo o tempo de execução e a fadiga imposta aos membros de um painel em avaliações sensoriais;

d) Validar o método proposto utilizando dados reais;

e) Comparar diferentes abordagens de análise sensorial voltadas ao desenvolvimento de novos produtos.

1.2 JUSTIFICATIVA DO TEMA E DOS OBJETIVOS

Aplicações sobre seleção de variáveis reportadas na literatura estão inseridas, em sua maioria, na área de Engenharia da Qualidade e abordam o problema sob a perspectiva do controle da qualidade. No contexto de Engenharia de Produção, estudos sobre o tema se justificam pelos seguintes aspectos: (*i*) um modelo composto por um grande número de variáveis pode apresentar uma alta aderência aos dados, mas não necessariamente um alto desempenho em termos de previsão e de classificação (devido à inclusão de variáveis ruidosas), conforme apontado por Guyson e Elisseeff (2003); (*ii*) a identificação de variáveis importantes baseadas no conhecimento empírico de engenheiros e operadores frequentemente leva a resultados tendenciosos; (*iii*) variáveis que inicialmente pareciam desempenhar um papel importante podem ter parecido relevantes devido a fatores secundários, tais como o mau funcionamento de sensores e outros dispositivos; desta forma, identificar variáveis realmente importantes é altamente desejável; e (*iv*) observância ao princípio de parcimônia, já que modelos destinados à caracterização de processos deveriam ser constituídos de poucas variáveis relevantes, de forma a facilitar a coleta e análise dos dados.

A qualidade vista pela Ciência de Alimentos é composta pelas características que diferenciam unidades individuais de um produto, sendo importante a determinação do grau de aceitabilidade pelo consumidor. O conceito de qualidade de um produto se refere àqueles atributos que o consumidor, consciente ou inconscientemente, estima que o produto deva possuir. Dessa forma, são considerados os atributos físicos, sensoriais, que devem estar associados para melhor entendimento das transformações que afetam ou não a qualidade do produto. Uma das formas de analisar a qualidade de um produto através de seus atributos é utilizando a análise sensorial, a qual compreende um conjunto de técnicas para medir atributos sensoriais de produtos a partir de respostas humanas (PIGGOTT *et al.*, 1998).

Nesse sentido, de forma específica na análise sensorial, este tema é relevante. A redução do número de variáveis que caracterizam e classificam um determinado produto resulta em: (*i*) redução do tempo necessário para avaliação dos produtos no painel sensorial, (*ii*) redução da fadiga dos provadores, uma vez que não terão que avaliar um elevado número de atributos; (iii) maior eficácia na especificação dos produtos, levando-os a maior qualidade final; e (*iv*) redução de custos devido à redução do tempo de avaliação sensorial pelos provadores.

A regressão por mínimos quadrados parciais (PLS, ou *Partial Least Squares*) tem sido amplamente utilizada na seleção de variáveis relevantes na previsão do desempenho de produtos, conforme reportado por Sarabia *et al.* (2001). Esse tipo de regressão foi concebido para apresentar um bom desempenho em bases de dados caracterizadas por (*i*) um grande número de variáveis, (*ii*) estrutura significativa de correlação, (*iii*) observações faltantes, e (*iv*) maior número de variáveis do que de observações (para detalhes, ver Kettaneh *et al.*, 2005). Aplicações de PLS na seleção de variáveis para fins de previsão são apresentadas por Wang *et al.* (2006) na indústria petroquímica, Wold *et al.* (2001) na indústria de reciclagem de papel, Xu *et al.* (2007) no tratamento de dados de emissões infravermelhas e Zhai *et al.* (2006) em dados quimiométricos. Outras aplicações relevantes são reportadas por Gauchi e Chagnon (2001) e Lazraq *et al.* (2003) em diferentes processos químicos.

Na indústria de alimentos, em particular no tratamento de dados sensoriais, a aplicação direta de PLS ou de técnicas baseadas neste tipo de regressão foi reportada por Kreutzmann *et al.* (2008) na caracterização de cenouras, e Granitto *et al.* (2007) na caracterização de diferentes tipos de queijos. Para fins de previsão com vistas ao controle da qualidade de processos e classificação de produtos, destacam-se um conjunto de aplicações na indústria vinícola, reportadas por Urtubia *et al.* (2007), e Capron *et al.* (2007). Tais aplicações na indústria alimentícia, entretanto, tinham por objetivo a construção de modelos de regressão a partir de técnicas de redução dimensional de dados; a exclusão de variáveis através da seleção daquelas relevantes na caracterização dos processos não foi contemplada nos estudos. A exceção é o trabalho de Sahmer e Qannari (2008), cujo objetivo é a identificação de um subconjunto de atributos sensoriais relevantes em dados de perfis sensoriais. Para tanto, os autores fazem uso da Análise de Componentes Principais (*Principal Componentes Analysis* ou PCA) em conjunto com análise de *clusters*. Apesar de identificarem variáveis relevantes de um grupo de candidatas, o método desenvolvido pelos autores não dispensa a coleta de dados sobre variáveis consideradas não-relevantes, no que se diferencia da proposta apresentada nesta tese.

Um segundo grupo de trabalhos aborda o tema da seleção de variáveis com vistas à classificação de produtos e processos; entre eles destacam-se Ortega (2000), Mallet *et al.* (1998), Piramuthu (2004) e Liu e Yu (2005). Muitas dessas aplicações focam na melhoria da capacidade de classificação de algoritmos de mineração de dados existentes na literatura, e na redução do tempo de processamento demandado por esses algoritmos usando um subconjunto de variáveis. Um levantamento completo das abordagens para seleção de variáveis baseadas

em mineração de dados é apresentado por Liu e Yu (2005), ao passo que uma panorâmica de aplicações industriais é apresentada por Mallet *et al.* (1998).

Um comparativo de algoritmos para seleção de variáveis é apresentado por Kudo e Sklansky (2000), onde o melhor subconjunto de variáveis é identificado combinando técnicas de classificação do tipo *leave-one-out* (deixar-uma-de-fora) e k-*nearest neighbor* (*k* vizinhos mais próximos). Além disso, Krzanowski (1995) utiliza análise discriminante para selecionar variáveis de um grupo que combina variáveis contínuas e discretas, ao passo que Pacheco *et al.* (2006) combinam abordagens de seleção de variáveis com análise discriminante (AD). Outras abordagens tratando da seleção de variáveis com objetivos preditivos incluem o uso de algoritmos genéticos (GUALDRON *et al.* 2007), métodos Bayesianos (GEORGE, 2000) e o método de regressão de Lasso (*Least Absolute Shrinkage and Selection Operator*) (FU, 1998).

Diante do exposto, pode-se observar a utilização de técnicas de análise multivariada em diferentes aspectos quanto à classificação e predição de produtos e processos. No entanto, estudos que conduzem ao desenvolvimento de um procedimento eficaz para identificar e selecionar os atributos relevantes na descrição e caracterização de um produto alimentício não foram identificados na literatura científica pesquisada.

1.3 DELINEAMENTO DO ESTUDO

Uma vez definidos os objetivos deste trabalho e apresentada a justificativa da importância desta pesquisa, é necessário estabelecer o delineamento do estudo pelo qual esses objetivos serão alcançados, considerando o método de pesquisa e o método de trabalho que serão utilizados.

1.3.1 Método de Pesquisa

A pesquisa realizada por este trabalho segue uma abordagem quantitativa. O ato de mensurar variáveis é a característica mais marcante da abordagem quantitativa. Dentro desta abordagem, o pesquisador deve capturar evidências da pesquisa por meio da mensuração das variáveis, com pequena ou nenhuma interferência nas variáveis de pesquisa (MIGUEL, 2010).

O desenvolvimento dos artigos apresentados segue uma das formas mais clássicas do método científico, a metodologia hipotética-dedutiva, que parte da percepção de uma lacuna nos conhecimentos acerca do qual se formulam hipóteses originadas de problemas

teóricos/práticos existentes, que devem ser submetidas à verificação com propósito de serem corroboradas (LAKATOS; MARCONI, 2005).

Em relação aos objetivos, a tese é classificada como pesquisa exploratória e aplicada. Segundo Gil (2002), as pesquisas exploratórias têm como objetivo proporcionar maior familiaridade com o problema, a fim de deixá-lo mais explícito ou a construir hipóteses, ou ainda, o aprimoramento de ideias ou a descoberta de intuições. A pesquisa aplicada, de acordo com Cervo e Bervian (2002), gera conhecimentos aplicados na prática com o intuito de solucionar problemas concretos.

1.3.2 Método de trabalho

O desenvolvimento do trabalho e execução das atividades a fim de alcançar os objetivos propostos ocorre através de cinco etapas, que são apresentadas em formato de artigos. Os artigos representam os meios para atingir o objetivo geral da tese. A estrutura do trabalho, com os artigos, seus objetivos e métodos, é apresentada na Figura 1.1.

Estudos	Objetivos	Questões de Pesquisa	Revisão Teórica	Método de Pesquisa
Artigo 1	Propor um método multivariado para seleção de atributos em painéis sensoriais com avaliações descritivas das amostras	Quais os principais atributos que conduzem à satisfatória acurácia de classificação das amostras em formulações?	1. Seleção de atributos em análise sensorial. 2. Análise de Componentes Principais e Análise Discriminante.	Pesquisa quantitativa: 1. Análise do método proposto 2. Validação em estudo de caso
Artigo 2	Desenvolver um método para identificar atributos que melhor discriminam amostras bem como painelistas que fornecem avaliações consistentes.	Quais os atributos que melhor discriminam amostras e qual o grupo de painelistas que fornecem avaliações consistentes?	1. Seleção de variáveis 2.Ferramentas analíticas: *k*-vizinhos mais próximos (KNN) e Análise de Pareto Ótimo.	Pesquisa quantitativa: 1. Análise do método proposto 2. Validação em estudo de caso
Artigo 3	Comparar métodos de análise multivariada, propondo a PLSDA como alternativa para discriminar produtos, avaliar julgadores e selecionar atributos .	Há características no PLSDA que o destacam na seleção de dados sensoriais ?	1. Ferramentas estatísticas para análise de perfis sensoriais. 2.PLSDA aplicado a perfis sensoriais convencionais. 3. Elipses de confiança.	Pesquisa quantitativa: 1. Análise do método proposto 2. Validação em estudo de caso
Artigo 4	Desenvolver um método para selecionar variáveis em dados colineares através de metodologia simples e comparar com um método descrito na literatura.	Quais as variáveis que melhor caracterizam os produtos? Qual o desempenho deste comparado com o método já descrito?	1. Métodos para seleçao de variáveis com ênfase na regressão PLS.	Pesquisa quantitativa: 1. Análise do método proposto 2. Validação em estudo de caso
Artigo 5	Comparar dois métodos para coleta e análise de dados sensoriais com vistas a dar suporte ao processo de desenvolvimento de um novo produto	Quais são as principais diferenças nos resultados obtidos a partir dos métodos considerados?	1. Análise sensorial 2. Métodos de análise sensorial 3. Grupos focados	Pesquisa qualitativa: 1. Estudo de caso em empresa no na área de desenvolvimento de produtos.

Figura 1.1: Estrutura das etapas da pesquisa desenvolvida

O Artigo 1 – "Seleção de Atributos em Avaliações Sensoriais Descritivas" – descreve o desenvolvimento de um estudo relacionado ao tema da pesquisa. Inicialmente, considera abordagens sobre seleção de atributos em análise sensorial evidenciadas na literatura com vistas à: (*i*) previsão das propriedades sensoriais e caracterização de produtos; (*ii*) identificação de segmentos de consumidores; e (*iii*) seleção de atributos em avaliações descritivas. Em seguida, descreve duas técnicas de análise multivariada de dados, Análise de Componentes Principais (PCA) e Análise Discriminante (AD), utilizadas no método proposto. Por fim, apresenta as etapas que levam à seleção dos atributos que maximizam a diferença entre as formulações avaliadas pelo painel sensorial.

O método proposto é composto por três fases principais, que conciliam as técnicas de análise multivariada citadas acima. O método é aplicado a um banco de dados de análise sensorial descritiva, através do método *Spectrum*, avaliando formulações de uma ração militar. Como resultado, chega-se a uma importante redução do número de atributos, mantendo o nível de acurácia similar à máxima possível, obtida quando utilizado o conjunto completo de atributos.

O Artigo 2 – "Método baseado na mineração de dados para identificação de atributos discriminantes em perfis sensoriais" – propõe o uso combinado de métodos de projeção multivariada e de mineração de dados para selecionar, simultaneamente, um subconjunto de julgadores consistentes e atributos discriminantes em painel sensorial com fins de descrição e discriminação dos produtos. A avaliação do método proposto foi realizada através de estudo de caso. Para tanto, utilizaram-se dados oriundos da avaliação sensorial de produtos militares constituídos de cubos de carne ao molho. Oito formulações diferentes foram avaliadas por um painel sensorial com nove membros em relação a vinte e quatro atributos.

A medida de consistência dos julgadores utilizada foi um índice associado à média ponderada da configuração, proposto por Leduphin *et al*. (2006). A seleção dos atributos discriminantes é implementada utilizando Análise de Componentes Principais (PCA) e a técnica de classificação dos *k*-vizinhos mais próximos (KNN), em conjunto com a análise do Pareto Ótimo (PO). A utilização deste conjunto de ferramentas mostrou-se adequada, permitindo uma redução significativa no número de atributos sem perda significativa na acurácia de classificação das amostras. Uma versão em inglês do artigo é apresentada no Apêndice da tese.

O Artigo 3 – "PLS discriminant analysis applied to conventional sensory profiling data" – aborda o principal tema de pesquisa através do uso da Análise Discriminante PLS (PLS-DA) como meio para a seleção de variáveis que diferencie produtos. Inicialmente, os dados foram pré-tratados a fim de eliminar fontes de variações entre os julgadores. Em seguida, analisaram-se alguns métodos descritos na literatura com foco no PLS-DA, apresentando suas vantagens em relação aos demais.

O PLS-DA fornece ferramentas estatísticas para avaliar tanto a concordância entre julgadores e a discriminação entre os produtos por meio da variância total, quanto a importância relativa das variáveis, permitindo a seleção de um subconjunto de atributos relevantes do conjunto completo de atributos, por meio do índice VIP (*Variable Importance in the Projection*).

O método PLS-DA foi comparado com a Análise de Componentes Principais (PCA) e Análise Canônica (CVA), e os resultados ilustrados através de um estudo de caso. Em particular, a estabilidade dos vários métodos foi investigada utilizando a reamostragem de julgadores (técnica de *bootstrap*) e elipses de confiança. Como resultado, PLS-DA proporcionou uma melhor estabilidade e melhor discriminação de produtos do que os métodos concorrentes.

O Artigo 4 – "Método de seleção de variáveis para minimização da variância de predição" – apresenta a seleção de variáveis em dados altamente multicolineares. O modelo propõe a seleção das variáveis através de critério de maximização da covariância. Este critério tem semelhança e propriedades análogas com a regressão PLS e é comparado com o método sugerido por Roger *et al.* (2011), denominado CovSel. Em um primeiro momento, uma revisão de trabalhos que relatam o uso de técnicas multivariadas para seleção de variáveis é descrita. Posteriormente, descreve-se a sequência de passos para implementação do método proposto.

Contrapondo-se ao CovSel, que considera somente a relação entre variáveis independentes (de predição) e a variável dependente (de resposta), o método proposto prevê a avaliação da correlação entre variáveis independentes (de predição) e destas com a variável dependente (de resposta). Os métodos foram avaliados em dois bancos de dados de espectro de infravermelho. Ambos os métodos apresentaram redução significativa no número de

variáveis; no entanto, o método proposto obteve uma redução maior de variáveis e índices de desempenho mais satisfatórios.

O Artigo 5 – "Comparação de diferentes abordagens na avaliação sensorial e desenvolvimento de produtos alimentícios" – contempla um estudo relativo à análise sensorial inserida no contexto de desenvolvimento de novos produtos. Aborda a comparação entre um método tradicional, o Teste de Ordenação Individual, e uma abordagem qualitativa baseada em Grupos Focados, através do estudo de caso de uma empresa produtora de achocolatado em pó. O uso de ambos os métodos conduziu a resultados similares; porém, o grupo focado se sobressai pela riqueza das informações oriundas das discussões em grupo, as quais servem como um importante subsídio para possíveis adequações de um determinado produto.

1.4 DELIMITAÇÕES DO ESTUDO

O presente trabalho incide sobre os aspectos de seleção, classificação, discriminação e predição de variáveis através de técnicas multivariadas de dados. Os métodos apresentados são voltados para aplicação em segmentos do setor alimentício e em quimiometria. No tocante a abrangência, a pesquisa concentra-se na área da análise sensorial descritiva e na análise de dados altamente multicolineares, como ocorridos em espectros de infravermelho.

Os modelos desenvolvidos envolvem dois tipos de banco de dados: os que contêm somente variáveis independentes, utilizados, sobretudo para estudos de classificação, e aqueles com uma variável dependente (variável de resposta). Estudos de bancos de dados com múltiplas variáveis dependentes não são contemplados nesta tese.

Além disso, o estudo está focado somente na análise relacionada a produtos e não será abordada, portanto, a análise de processos.

1.5 ESTRUTURA DA TESE

Esta proposta de tese está organizada em sete capítulos principais. Neste primeiro capítulo foram apresentados a contextualização do trabalho e os objetivos, justificando a importância desta pesquisa desde o ponto de vista acadêmico e prático. Este capítulo também apresentou o método de trabalho, a estrutura e as delimitações do estudo. Os capítulos posteriores, de dois a seis, apresentam os artigos contendo os desenvolvimentos propostos,

conforme a estrutura apresentada anteriormente na Figura 1. O sétimo capítulo apresenta as conclusões da tese e futuras pesquisas a serem desenvolvidas a partir dos resultados já obtidos.

1.6 REFERÊNCIAS

ANZALDÚA-MORALES, A. A., **La evaluación sensorial de los alimentos en la teoria e la práctica***,* Zaragoza: Editorial Acribia S.A., 1994.

ANZANELLO, M. J., FOGLIATTO, F. S., ROSSINI, K., Data mining-based method for identifying discriminant atributes in sensory profiling, **Food Quality and Preference,** v.22, p.139-148, 2011.

BAUMANN, K., ALBERTO, H., KORFF M., A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part I. Search algorithm, theory and simulations**, Journal of Chemometrics**, v.16, p.339-350, 2002.

CAPRON, X., SMEYERS-VERBEKE, J., MASSART, D. Multivariate determination of the geographical origin of wines from four different countries. **Food Chemistry**, v.101, p.1585–1597, 2007.

CARBONELL, L.; IZQUIERDO, L.; CARBONELL, I. Sensory analysis of Spanish mandarin juices. Selection of attributes and panel performance, **Food Quality and Preference***,* v.18, p.329–341, 2007.

CERVO, A.L.; BERVIAN, P.A. **Metodologia científica**. 5ª ed. São Paulo: Prentice Hall, 2002, 242 P.

DUTCOSKY, S. D. **Análise Sensorial de Alimentos.** 2ª Edição. Curitiba: Editora Champagnat - Coleção Exatas 4, 2007. 239 p.

FERRAND, M., HUQUET, B. BARBEY,S. BARILLET,F. F. FAUCON, LARROQUE, H. LERAY, O. TROMMENSCHLAGER, J.M., BROCHARD M., Determination of fatty acid profile in cow's milk using mid-infrared spectrometry: Interest of applying a variable selection by genetic algorithms before a PLS regression, **Chemometrics and Intelligent Laboratory System**s, 2010. In press.

GAUCHI, J., CHAGNON, P. Comparison of selection methods of exploratory variables in PLS regression with application to manufacturing process data. **Chemometrics and Intelligent Laboratory System**s, v.58, p.171–193, 2001.

GIL, A.C. **Como elaborar projetos de pesquisa**. 4ª ed. São Paulo: ATLAS, 2002, 175p.

GIL, A.C. **Métodos e Técnicas de Pesquisa Social**. 6ed. São Paulo: Atlas, 2008, 200p.

GRANITTO, P.M.; GASPERI, F.; BIASIOLI, F.; TRAINOTTI, E.; FURLANELLO, C. Modern data mining tools in descriptive sensory analysis: A case study with a Random forest approach, **Food Quality and Preference**, v.18, p.681–689, 2007.

GUALDRON, O., LIOBET, E., BREZMES,J., VILANOVA, X., CORREIG, X. Fast variable selection for gas sensing applications, **Proceedings of IEEE Sensors**, p.892-895, 2004.

GUYON, I., ELISSEEFF, A., An introduction to variable and feature selection. **Journal of Machine Learning Research**, v.3, p.1157–1182, 2003.

KETTANEH, N.; BERGLUND, A.; WOLD, S. PCA and PLS in very large datasets. **Computational Statistics & Data Analysis**, v.48, p.69-85, 2005.

KREUTZMANN, S.; SVENSSON, V.T.; THYBO, A.K.; BRO, R.; PETERSEN, M.A. Prediction of sensory quality in raw carrots (Daucus carota L.) using multi-block LS-ParPLS, **Food Quality and Preference**, v.19, p.609–617, 2008.

KRZANOWSKI, W. Selection of variables, and assessment of their performance, in mixed-variable discriminant analysis. **Computational Statistics and Data Analysis**, v.19, p.419-431, 1995.

KUDO, M., SKLANSKY, J. Comparison of algorithms that select features for pattern classifiers. **Pattern Recognition**, v.33, p.25–41, 2000.

LAKATOS, E.M.; MARCONI, M.A. **Fundamentos de metodologia científica**. 6ª ed. São Paulo: ATLAS, 2005, 315P.

LAWELESS, H. T.; HEYMANN, H. **Sensory Evaluation of Food: Principles and Practices.** New York: Chapman & Hall, 1998. 819p.

LAZRAQ, A., CLEROUX, R., GAUCHI, J. Selecting both latent and exploratory variables in

the PLS1 regression model, **Chemometrics and Intelligent Laboratory System**s, v.66, p.117-126, 2003.

LEARDI, R. SEASHOLTZ, M. PELL, R. Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data, **Analytica Chimica Acta,** v.461, p.189-200, 2002.

LEDAUPHIN, S.; POMMERET, D.; QANNARI, M. Application of hidden Markov model to products shelf lives. **Food Quality and Preference**, v.19, p.156–161, 2008.

LIU, H., YU, L.. Toward integrating feature selection algorithms for classification and clustering. **IEEE Transactions on Knowledge and Data Engineering**, v.17(4), p.491–502, 2005.

MALLET, Y., DE VEL, O., COOMANS, D. Integrated feature extraction using adaptive wavelets. In: H. Liu & H. Motoda, **Feature extraction, construction and selection: A Data mining perspective**, p.175–189, 1998.

MIGUEL, P.A.C.; FLEURY, A.; MELLO, C.H.P.; NAKANO, D.N.; TURRIONI, J.B.; HO, L.L.; MORABITO, R. MARTINS, R. A. **Metodologia de pesquisa em engenharia de produção e gestão de operações.** Rio de Janeiro: Elsevier, 226p, 2010.

MURRAY, J. M., DELAHUNTY, C. M., & BAXTER, I. A., Descriptive sensory analysis: Past, present and future. **Food Research International**, v.34(6), p.461–471, 2001.

PIGGOTT, J. R.; SIMPSON, S. J.; WILLIAMS, S. A. R. Sensory analysis. **International Journal of Food Science and Technology**, v.33, p.7-18, 1998.

ROGER, J.M., PALAGOS, B., BERTRAND, D., FERNANDEZ-AHUMADA, E., CovSel: Varible selection for highly multivariate and multi-response calibration Application to IR spectroscopy, **Chemometrics and Intelligent Laboratory Systems**, v.106, p.216-223, 2011.

ROY, P., ROY, K., On Some Aspects of Variable Selection for Partial Least Squares Regression Models, QSAR, **Combinatorial Science,** v.27, p.302-313, 2008.

SAHMER, K.; QANNARI, E.M. Procedures for the selection of a subset of attributes in sensory profiling. **Food Quality and Preference**, v.19, p.141–145, 2008.

SARABIA, L., ORTIZ, M., SANCHEZ, A., Dimension wise selection in partial least squares regression a bootstrap estimated signal-noise relation to weight the loadings, in: PLS and Related Methods, Proc. PLS'01 **International Symposium**, CISIA-CERESTA Editeur, Paris, (2001) 327-339.

URTUBIA, A., PERREZ-CORREA, J., SOTO, A., & PSZCZOLKOWSKI, P. Using data mining techniques to predict industrial wine problem fermentation. **Food Control**, v.18, p.1512–1517, 2007.

WANG, D., SRINIVASAN, R., LIU, J., GURU, P., LEONG, K. Data-driven soft sensor approach for quality prediction in a refinery process, **Proceedings of IEEE International Conference Ind. Inf.** 2006.

WOLD, S., TRYGG, J., BERGLUND, A., ANTTI, H., Some recent developments in PLS modeling. **Chemometrics and Intelligent Laboratory Systems**, v.58, p.131-150, 2001b.

XU, L., JIANG, J., WU, H., SHEN, G., YU, R., Variable-weighted PLS, **Chemometrics and Intelligent Laboratory Systems,** v.85, p.140-143, 2007.

ZHAI, H., CHEN, X. HU, A., A new approach for the identification of important variables, **Chemometrics and Intelligent Laboratory Systems**, v.80, p.130-135, 2006.

2 ARTIGO 1 – Seleção de Atributos em Avaliações Sensoriais Descritivas

Karina Rossini

Michel J. Anzanello

Flávio S. Fogliatto

Artigo enviado para publicação na revista Produção (ABEPRO); no prelo.

Resumo

A seleção dos atributos a serem avaliados em uma análise sensorial é uma importante etapa no planejamento de painéis sensoriais. O processo de seleção visa reduzir a lista de atributos a serem apresentados aos julgadores, evitando assim a imposição de fadiga aos membros do painel, porém mantendo atributos significativos na caracterização das amostras avaliadas. Este artigo apresenta um método multivariado para seleção de atributos em painéis sensoriais baseados em avaliações descritivas das amostras, tais como os métodos QDA (*Quantitative Descriptive Analysis*) e Spectrum. O método proposto utiliza Análise de Componentes Principais para identificação dos atributos mais relevantes, e então aplica Análise Discriminante para classificação das amostras em formulações distintas. O método é aplicado em um estudo de caso onde cubos de carne com molho, utilizados em ração de combate pelo exército norte-americano, são caracterizados em painel sensorial utilizando o método QDA. O método proposto reduz significativamente o número de atributos a serem avaliados pelo painel sensorial e conduz à satisfatória acurácia de classificação das amostras em formulações.

Palavras-Chave: Seleção de atributos; Análise sensorial; Análise multivariada

Abstract

Selection of attributes from a group of candidates to be assessed through sensory analysis is an important step when planning sensory panels. When selecting attributes it is desirable to reduce the list of those to be presented to panelists avoiding fatigue, however keeping attributes that are relevant in the sensory characterization of samples. In this paper we present a multivariate method for attribute selection in descriptive sensory panels, such as those used in the QDA (Quantitative Descriptive Analysis) e Spectrum protocols. The

proposed method is implemented using Principal Component Analysis and Descriptive Analysis, and is evaluated in a case study where beef cubes in stew, used as combat ration by the American Army, are characterized in sensory panels using the Spectrum method. The method significantly reduced the number of attributes to be considered in sensory panels, while yielding satisfactory accuracy in the classification of samples.

Keywords: Attribute selection; Sensory evaluation; Multivariate analysis

2.1 INTRODUÇÃO

Na indústria de alimentos, a análise sensorial é elemento chave para identificar as expectativas dos consumidores (LEDAUPHIN *et al.*, 2008). A análise sensorial compreende um conjunto de técnicas para medir precisamente atributos sensoriais de produtos a partir de respostas humanas. Tais técnicas utilizam princípios oriundos da ciência de alimentos, fisiologia, psicologia e estatística, fornecendo respostas objetivas para as propriedades de alimentos, conforme percebidas pelos cinco sentidos (PIGGOTT *et al.*, 1998).

Informações obtidas através de avaliações sensoriais podem ser utilizadas por empresas como suporte técnico para pesquisa, industrialização, marketing e controle de qualidade dos itens produzidos, qualificando decisões técnicas e administrativas. Na perspectiva do consumidor, a avaliação sensorial em produtos industriais assegura que os mesmos cheguem ao mercado com as características desejadas (DUTCOSKI, 1996; LAWLESS; HEYMANN, 1998).

A avaliação sensorial é realizada de acordo com diferentes testes que dependem da sua finalidade (ANZALDÚA-MORALES, 1994). Os métodos descritivos, dentre eles a análise descritiva quantitativa (ADQ), estão entre as ferramentas mais elaboradas da ciência sensorial e envolvem a detecção (discriminação) e descrição tanto de componentes sensoriais qualitativos quanto quantitativos por um painel treinado de julgadores. Os julgadores sensoriais devem quantificar os aspectos dos produtos de maneira a facilitar a descrição da percepção dos seus atributos. A principal vantagem da análise descritiva está em sua habilidade de permitir que seja determinada uma relação entre medidas sensoriais descritivas e a instrumental ou de preferência do consumidor (MURRAY *et al.*, 2001). Limitações comuns nesses testes referem-se ao número de amostras, número de julgadores ou quantidade de atributos a serem analisados. No que tange a esse último aspecto, Sahmer e Qannari (2008) apontam que a seleção de atributos tem sido foco de investigações em diversas áreas

industriais. Em particular, no perfil sensorial descritivo, estratégias de seleção podem ser utilizadas para reduzir o conjunto de atributos, selecionando aqueles relevantes e não redundantes. Tal seleção reduz o tempo de execução, a fadiga imposta aos membros do painel, e os custos da avaliação.

Um atributo significativo em uma investigação sensorial é aquele cuja avaliação apresenta diferenças sistemáticas e significativas entre as amostras investigadas, tal que seja possível relacionar o nível do atributo (através de seu valor medido) com características das amostras (por exemplo, a presença ou ausência de um ingrediente). Atributos que apresentam tal comportamento poderiam, por exemplo, ser utilizados como variáveis de resposta em modelos de regressão, os quais permitiriam a otimização do produto investigado no painel sensorial. Neste artigo, propõe-se um método para identificação de tais atributos usando dados obtidos em uma análise sensorial preliminar usando métodos descritivos.

O método proposto utiliza ferramentas de análise multivariada de dados, conciliando a Análise de Componentes Principais (PCA) e a Análise Discriminante (AD) como meio para redução do número de atributos (variáveis) na avaliação sensorial. Os pesos gerados pela PCA são transformados em índices de importância dos atributos. Esses índices são vinculados à AD, a qual classifica as amostras em diferentes classes de formulação, e uma medida de desempenho da classificação é calculada. Na sequência, o atributo apontado como menos relevante pelo índice de importância é eliminado e uma nova classificação é realizada, acompanhada por uma nova medição de desempenho. Esse procedimento iterativo é finalizado ao atingir-se um número limite de atributos remanescentes.

Na PCA, as variáveis são reescritas como combinações lineares, as quais são denominadas componentes principais. Cada componente descreve uma porção da variabilidade presente nas variáveis originais e sua interpretação é geralmente baseada na magnitude dos pesos associados às variáveis. Apesar do número de componentes principais resultantes de uma PCA sobre um conjunto de variáveis ser igual ao número de variáveis analisadas, usualmente é possível obter uma boa representação dos dados utilizando um número reduzido de componentes (RENCHER, 1995). A AD, por sua vez, é uma técnica de classificação e discriminação de amostras, que permite alocar novas observações a grupos pré-determinados. Na utilização da AD, um grupo de observações cujos membros já estão identificados é utilizado para estimar pesos (ou parâmetros) de uma função discriminante

conforme alguns critérios, tal como a minimização de erros de classificação (SUEYOSHI e GOTO, 2009).

O presente artigo traz uma importante contribuição para a área da Sensometria: o desenvolvimento de procedimentos de seleção de atributos, usuais em estudos de classificação, aplicados a problemas de análise sensorial. Os métodos QDA e Spectrum, de ampla utilização prática, estão baseados na caracterização plena das amostras através de um grande número de atributos, muitos dos quais não contribuem na diferenciação das amostras investigadas. O método aqui proposto permite identificar tais atributos utilizando dados de uma análise preliminar, eliminando-os da coleta posterior de dados, envolvendo um maior número de julgadores. Otimiza-se, assim, a coleta de dados sensoriais, usualmente de alto custo no desenvolvimento de produtos industriais.

O presente trabalho é composto de cinco seções. Além da introdução, é apresentado um referencial teórico sobre seleção de variáveis, seguido pela descrição do método proposto e, posteriormente, apresenta-se o estudo aplicado, finalizando com as conclusões do mesmo.

2.2 REFERENCIAL TEÓRICO

As seções seguintes apresentam abordagens para a identificação dos atributos mais relevantes em análise sensorial, além de descrever os fundamentos das técnicas multivariadas utilizadas no método proposto (Análise de Componentes Principais e Análise Discriminante).

2.2.1 Seleção de atributos em análises sensoriais

Técnicas analíticas multivariadas têm sido amplamente utilizadas em diversas áreas do conhecimento, motivadas por variadas aplicações. Na indústria de alimentos, em particular no tratamento de dados sensoriais, ressalta-se o envolvimento das mesmas para fins de (*i*) previsão das propriedades sensoriais e caracterização de produtos; (*ii*) seleção e identificação de segmentos de consumidores; e (*iii*) seleção de atributos em análises sensoriais descritivas.

Com vistas à previsão das propriedades sensoriais e caracterização de produtos, Johansen *et al.* (2008), em pesquisa com iogurtes desnatados e *cream cheese*, utilizaram a Regressão por Mínimos Quadrados Parciais (PLSR, ou *partial least squares regression*) para relacionar os resultados da análise sensorial descritiva a imagens digitais das superfícies das amostras. Karoui *et al.* (2006) fizeram uso de PLSR e Análise de Correlação Canônica (CCA,

ou *canonical correlation analysis*) para comparar os resultados da análise sensorial com os obtidos através do Infra-Vermelho Próximo (NIR, ou *near infra red*) em queijos, ao passo que Esteban-Díez *et al.* (2004) utilizaram estas mesmas técnicas para investigar as relações entre determinados atributos sensoriais do café expresso e dados espectrais de amostras de café torrado. Por fim, Kreutzmann *et al.* (2008) utilizaram uma abordagem multi-bloco (LS-ParPLS) para a caracterização de cenouras.

Uma sistemática para identificação de segmentos de consumidores foi proposta por Sahmer *et al.* (2006) através da combinação de análise de *cluster* e algoritmo hierárquico em estudos com café. Heenan *et al.* (2008) avaliaram três segmentos de consumidores quanto à percepção acerca de pães frescos através da Análise de Componentes Principais (PCA) e análise de *cluster*, utilizando também PLSR para investigar a relação entre as percepções dos consumidores para cada segmento. Carbonell *et al.* (2008) investigaram a segmentação de consumidores através do uso de um coeficiente de correlação entre consumidores e atributos sensoriais. Naquele trabalho, a segmentação, por meio da análise de *clusters*, fundamentou-se na semelhança entre os coeficientes de correlação dos consumidores.

A seleção de atributos em análises sensoriais descritivas, tema do presente artigo, é abordado por outro grupo de pesquisadores. Dijksterhuis *et al.* (2002) propuseram uma aplicação em análise sensorial para o método estatístico proposto em Krzanowski (1987): o estudo objetivou reduzir um grande grupo de variáveis determinantes da percepção da gordura no leite em subgrupos menores, utilizando PCA. Os autores utilizaram uma rotação *procrustes* nos subgrupos, sendo o subgrupo detentor da menor perda *procrustes* selecionado. Já Sahmer e Qannari (2008) compararam seis métodos distintos para identificação de um subconjunto de atributos sensoriais relevantes em dados descritivos oriundos de perfis sensoriais. Dois dos métodos testados baseiam-se em PCA, dois utilizam a análise generalizada de *procrustes* (GPA, ou *generalized procrustes analysis*) e dois envolvem análise de *clusters*. A análise visou reduzir o número total de atributos apresentados ao painel e preservar a estrutura multivariada dos dados. Os métodos utilizando análise de *clusters* obtiveram melhores resultados em conjuntos de dados reais e simulados. Porém, apesar de proporcionar a identificação dos atributos relevantes, o método desenvolvido pelos autores não dispensa a coleta de dados sobre atributos considerados não-relevantes.

Westada *et al.* (2003) apresentam uma metodologia para identificação de atributos significantes em modelos multivariados, baseada na aplicação de PCA em dados de análise

sensorial descritiva e de consumidores. O objetivo é interpretar fatores latentes que abrangem características como sabor, odor, aparência e textura em sorvetes e queijos, identificando produtos semelhantes ou diferentes, bem como os atributos que os diferenciam. O método foi baseado nas estimativas de incerteza provenientes da técnica de *cross-validation/Jack–knifing*, sendo o teste *t* de *student* utilizado para calcular a significância de cada variável para cada componente.

Carbonell *et al.* (2007) utilizaram a GPA para selecionar os atributos na avaliação de sucos natural e processado de tangerina. Peron (2000) identificou as diferenças entre variedades de batatas através de trinta atributos avaliados por 14 julgadores. Para tanto, selecionaram-se atributos de acordo com sua capacidade discriminativa utilizando ANOVA (Análise de Variância) e PCA. Por fim, o autor utilizou uma GPA normalizada para calcular o índice confiabilidade de cada julgador. Outro método para seleção de atributos discriminantes, desta vez aplicado a diferentes tipos de queijos, foi proposto por Granitto *et al.* (2007). Os autores compararam o método *random forests* (RFs) com análise discriminante linear (LDA, ou *Linear Discriminant Analysis*) e discriminante por mínimos quadrados parciais (DPLS, ou *Discriminant Partial Least Squares*). A função discriminante derivada do RFs atruibui pesos de importância para os atributos, sendo esses pesos responsáveis pela seleção de um número reduzido de atributos que garanta um satisfatório poder discriminatório da função.

2.2.2 Análise de Componentes Principais e Análise Discriminante

A Análise de Componentes Principais (PCA) tem o objetivo de reduzir a dimensionalidade de um conjunto de dados sem perda significativa de informações. Os componentes principais são combinações lineares dos dados originais. Através da PCA é possível substituir os dados originais por um número reduzido de componentes principais independentes entre si (JACKSON, 1980 e 1981).

Considere um conjunto de dados composto de realizações de *p* atributos sensoriais. É possível extrair desse conjunto de dados *p* componentes principais, sendo cada um uma combinação linear distinta dos *p* atributos originais. Cada componente principal (CP) captura uma direção de variabilidade do conjunto de dados originais. As direções capturadas por cada componente principal são ortogonais entre si (ou seja, garantindo a independência entre CPs).

A determinação dos componentes principais pode utilizar as informações na matriz de covariâncias ou de correlações associadas aos *p* atributos originais. A matriz de

covariâncias é recomendada quando a escala dos atributos analisados pode fornecer informações relevantes sobre a estrutura dos dados (RENCHER, 1995). No caso de atributos descritos por escalas oriundas de questionário ou painéis, geralmente contidas no mesmo intervalo de valores, a análise através da matriz de covariâncias e de correlação costuma gerar os mesmos resultados, a menos que haja efeito significativo de localização no uso da escala pelos diferentes julgadores. Nesse caso, recomenda-se a remoção de tais efeitos através de pré-tratamento dos dados (por exemplo, usando o método proposto em Ledauphin *et al.*, 2006) ou o uso da matriz de correlações na PCA.

Seja $\mathbf{\Sigma}$ a matriz de covariâncias, de dimensão ($p \times p$), associada à matriz de atributos $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p]$. A dimensão de $\mathbf{X}$ é ($n \times p$), ou seja, dispõe-se de n observações de cada atributo. O vetor $\mathbf{x}^t$ denota uma linha qualquer de $\mathbf{X}$. Os p autovalores de $\mathbf{\Sigma}$ são designados por λ_i, $i = 1,\ldots, p$, e os p autovetores designados por $\mathbf{w}_i$, $i = 1,\ldots, p$, com elementos dados por ($w_{i1},\ldots, w_{ip}$). Associado a cada autovalor λ_i existe um autovetor $\mathbf{w}_i$. Assim, os pares (λ_1, $\mathbf{w}_1$), (λ_2, $\mathbf{w}_2$), ..., (λ_p, $\mathbf{w}_p$) correspondem aos autovalores e autovetores de $\mathbf{\Sigma}$, com autovalores arranjados tal que $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_p$. O i-ésimo componente principal pode ser obtido pela expressão (SEBER, 1984):

$$Y_i = \mathbf{x}^t \mathbf{w}_i = w_{i1} X_1 + w_{i2} X_2 + \ldots + w_{ip} X_p \,. \qquad (1)$$

O número de componentes principais a ser retido (R) pode ser definido com base no percentual de variância explicado pelos componentes. Para tanto, somam-se os autovalores (arranjados do maior ao menor) até atingir-se um valor desejado de variância (RENCHER, 1995). O percentual ideal da variância explicada pelos PCs depende da natureza do problema analisado.

Na eq. (1), os elementos do autovetor $\mathbf{w}_i$ funcionam como pesos de importância de $X_1,\ldots, X_p$ na composição do i-ésimo CP, e são denominados cargas do componente. Em CPs obtidos a partir de atributos padronizados, a magnitude do peso associado ao atributo descreve sua importância relativa na composição dos CPs.

No restante desta seção, apresentam-se os fundamentos da Análise Discriminante (AD).

A AD é uma técnica de classificação de observações em classes distintas (JOBSON, 1996). Diferentemente das técnicas tradicionais de agrupamentos a AD possibilita a classificação de novos objetos nas classes já existentes sem que seja necessário um rearranjo de classes (MINGOTI, 2005). Além do propósito de classificação, a AD também pode ser utilizada para obter um melhor entendimento das relações entre as observações e os grupos a que pertencem.

A técnica é operacionalizada através da construção de uma relação entre uma variável dependente não-contínua (*Z*) e um conjunto de variáveis contínuas (*x's*). Essa relação é expressa através de uma função discriminante consistindo em uma combinação das variáveis contínuas (HAIR *et al*., 2006; ZANINE *et al*., 2008). Duas funções tradicionalmente geradas são as do tipo linear [aplicada quando as matrizes de covariâncias são similares entre as classes analisadas e exemplificada na equação (2)] ou do tipo quadrática, quando as matrizes de covariâncias entre as classes são distintas. Outras funções podem ser encontradas em *Hair* et al. (2006).

$$Z = b_0 + b_1 x_1 + ... + b_p x_p , \qquad (2)$$

onde b_0 é uma constante, b_i são os coeficientes de ponderação associados às variáveis independentes, e *Z* é o escore discriminante obtido para cada objeto na análise.

Por constituir-se em uma técnica supervisionada (Duda *et al*., 2001), os coeficientes b_i da AD são estimados com base em um banco de dados onde a classe de cada observação é conhecida *a priori*. Os coeficientes são determinados com base na média e na variância-covariância das variáveis pertencentes a cada classe, maximizando-se a diferença padronizada entre os escores médios provenientes de classes distintas. A classificação de observações futuras é feita através da determinação do seu escore *Z*, e posterior comparação com limites estabelecidos. Mais detalhes acerca da geração da função discriminante e obtenção dos coeficientes podem ser encontrados em Hair et al. (2006).

2.3 MÉTODO PROPOSTO

O método aqui proposto é baseado em uma sequência de etapas que utilizam, majoritariamente, duas técnicas de análise multivariada de dados: PCA e AD. Os pesos dos componentes principais da PCA geram subsídios para a identificação dos atributos que

apresentam maior variabilidade, servido como base para a geração de um índice de importância dos atributos. A AD, por sua vez, cumpre a função de categorizar as observações (amostras) nas classes de formulação com base nos atributos. Um processo iterativo de classificação via AD e eliminação dos atributos menos relevantes é balizado pelo índice de importância gerado para os atributos.

A aplicação do método pressupõe uma pré-coleta de dados sensoriais utilizando métodos descritivos, tais como o QDA (*Quantitative Descriptive Analysis*; ver Stone et al., 1974, e Stone e Sidel, 1992) e o método Spectrum (ver Meilgaard et al., 1999) . A principal característica desse tipo de método é a mensuração da intensidade de atributos em amostras utilizando escalas contínuas. Na pré-coleta, utiliza-se um número reduzido de julgadores e repetições. De posse dos dados da pré-coleta, aplicam-se as etapas detalhadas no restante desta seção.

O método fundamenta-se na seguinte lógica. Deseja-se identificar o conjunto de atributos que maximiza a diferença entre as formulações, aqui tratadas como classes, avaliadas no painel sensorial. Quanto maior a diferença entre classes, maior a chance da obtenção de modelos de regressão significativos, associando atributos a ingredientes testados em diferentes níveis nas formulações. Tais modelos viabilizam a otimização da formulação do produto analisado usando procedimentos de otimização de experimentos multiresposta (ver FOGLIATTO e ALBIN, 2001, entre outros). Desta forma, atributos a serem retidos são aqueles responsáveis por parcelas importantes da variabilidade observada entre classes. O resultado de um painel sensorial fornecido por um julgador é uma observação; nela, uma amostra, correspondendo a uma formulação, é analisada com relação à p atributos. Conhece-se, *a priori*, a qual classe pertence cada observação obtida na análise sensorial do produto, já que as mesmas estão vinculadas às formulações testadas; o objetivo aqui é identificar quais atributos melhor caracterizam essa alocação de observações em classes.

2.4 APLICAÇÃO DA PCA NO BANCO DE DADOS E GERAÇÃO DO ÍNDICE DE IMPORTÂNCIA DOS ATRIBUTOS

A PCA é aplicada na matriz de correlação dos atributos e tem por objetivo gerar pesos de importância para os mesmos. Tais pesos, dados pelo autovetor $\mathbf{w}_i$ associado ao i-ésimo componente principal, além de identificar atributos que apresentam maior

variabilidade, são utilizados na geração de um índice de importância para cada atributo (designado por IA_i).

O índice de importância do atributo *i* é estimado com base na magnitude dos pesos gerados pela PCA. Este índice é composto pela soma dos valores absolutos dos pesos sobre os *R* componentes retidos, conforme apresentado na equação (3).

$$IA_i = \sum_{j=1}^{R} | w_{ij} | \qquad i=1,\ldots,p \qquad (3)$$

Atributos com valores de *IA* elevados na equação (3) são considerados mais relevantes na análise, uma vez que apresentam maior variância associada. Segundo Duda et al. (2001), atributos com elevada variância são desejados em procedimentos de classificação/discriminação, visto que possibilitam separações mais precisas entre os grupos.

2.4.1 Eliminação dos atributos irrelevantes para classificação das amostras em classes de formulação

Neste passo, almeja-se identificar os atributos mais relevantes para classificação das amostras em classes de formulação. Para tanto, adota-se uma sistemática de eliminação dos atributos menos relevantes no formato *backward*, onde os atributos são sistematicamente eliminados do banco de dados após cada classificação.

O procedimento de classificação proposto aplica AD sobre os *p* atributos. O desempenho da classificação das amostras (observações) nas classes pré-definidas (formulações) é estimado através da acurácia. A acurácia é definida como a razão entre o número de classificações corretas e o número total de classificações efetuadas. É importante enfatizar que o procedimento de classificação não se apoiou nas tradicionais porções de treino (construção do modelo) e de teste (validação em novas observações) em decorrência do limitado número de repetições dentro de cada combinação de formulação e julgador. Logo, a acurácia de classificação foi estimada com base nas mesmas observações utilizadas para construção da função discriminante.

Na sequência, o atributo com o menor valor de *IA* é eliminado do banco de dados, e os *p*-1 atributos remanescentes são utilizados em uma nova classificação das amostras. A acurácia de classificação é novamente calculada. Esse procedimento iterativo é mantido até

atingir-se um número desejado de atributos remanescentes, sendo que dois atributos é o limite inferior recomendado.

Um gráfico associando acurácia e número de atributos retidos é gerado para monitorar o processo de eliminação. Esse gráfico auxilia na definição de um subconjunto apropriado de atributos, o qual é analisado em detalhes na etapa seguinte do método.

2.4.2 Análise do gráfico de acurácia e definição do melhor subconjunto de atributos

Essa etapa é composta por duas subetapas: (*i*) definição de um ponto de corte no gráfico de acurácia e seleção de um subconjunto de atributos, *SA*; e (*ii*) enumeração das possíveis combinações dos atributos em *SA* (incluindo um número limitado de atributos recentemente eliminados, *AE*), e seleção do subconjunto de atributos responsável pela máxima acurácia de classificação.

A definição de um ponto de corte no gráfico de acurácia é feita de forma subjetiva, e visa definir um número mínimo de atributos a ser considerado no procedimento de enumeração. Recomenda-se escolher um ponto que concilie um nível de acurácia satisfatório e um número reduzido de atributos retidos. Essa condição é verificada próxima a pontos onde a acurácia apresenta considerável redução com a eliminação de um atributo, conforme exemplo na Figura 1.

Na sequência, recomenda-se incluir alguns dos atributos recentemente eliminados (ou seja, pertencentes ao subconjunto *AE*) ao subconjunto *SA* (conforme Figura 2.1). Esse procedimento visa assegurar que atributos de importância intermediária sejam inseridos no procedimento de enumeração, podendo eventualmente gerar acurácias satisfatórias ao serem combinados aos atributos em *SA*. A não inclusão de elementos de *AE* em *SA* não compromete a eficiência do método, mas pode conduzir a acurácias inferiores. O número de atributos de *AE* é definido subjetivamente (recomenda-se de 3 a 5 atributos), sendo escolhidos os atributos com valor de *IA* imediatamente superior ao *IA* do último atributo incluído no subconjunto *SA*.

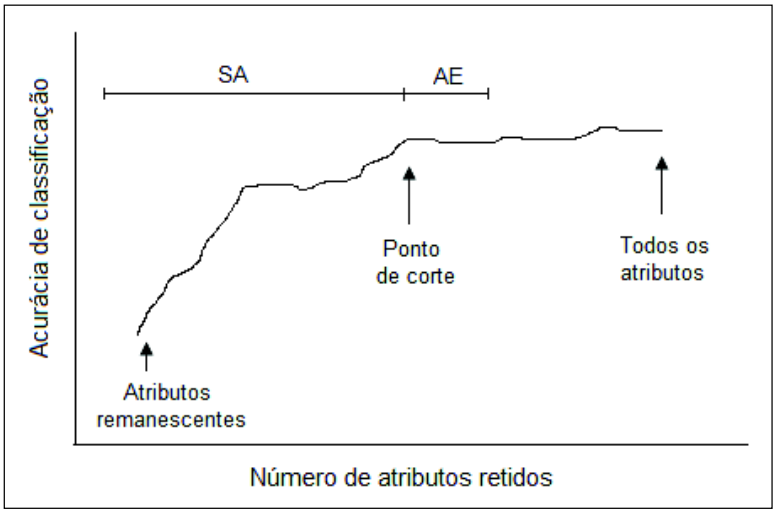


Figura 2.1: Ponto de corte e atributos retidos na análise

A subetapa (*ii*) consiste na enumeração total de combinações $\left[\binom{m}{n}\right.$, onde *m* refere-se aos atributos em *SA*+*AE* e *n* denota o possível número de atributos a serem mantidos, variando de {*SA*} a {*SA*+*AE*}, onde {ϕ} denota o número de elementos em ϕ]. Cada combinação tem sua acurácia de classificação avaliada, sendo que a combinação responsável pela máxima acurácia indica o melhor subconjunto de atributos.

Por fim, enfatiza-se que uma análise simplificada pode ser conduzida através da seleção direta dos atributos responsáveis pelo ponto de corte no gráfico de acurácia (subconjunto *SA*), eliminando-se assim a necessidade de enumeração.

2.5 ESTUDO DE CASO

Nesta seção, aplica-se o método proposto para seleção de atributos utilizando dados obtidos em uma avaliação sensorial realizada utilizando o método Spectrum de análise sensorial descritiva. Trata-se de um dos mais rigorosos métodos para treinamento e coleta de informações sensoriais de julgadores, desenvolvido por Gail Civille e apresentado em Meilgaard et al. (1999).

Julgadores candidatos a integrar o painel foram selecionados em duas etapas. A primeira etapa consistiu de uma entrevista telefônica, com o objetivo de identificar a disponibilidade do candidato e sua familiaridade com testes sensoriais. Candidatos

considerados aptos foram convidados a realizar um bateria completa de testes, nos quais atributos relacionados a sabor, textura e aparência foram investigados. De aproximadamente 25 candidatos, 8 foram selecionados ao término da fase preliminar de treinamento (3 meses). A segunda fase consistiu de duas sessões de treinamento semanais, de duração aproximada de 1,5 horas, durante mais 3 meses, ao cabo dos quais os julgadores foram considerados treinados.

Os painéis sensoriais buscavam a caracterização de uma ração militar constituída de cubos de carne ao molho embalados em *pouches* termoestáveis. Os oito julgadores analisaram 26 atributos de 8 formulações do produto em questão. Cada avaliação foi realizada em quadriplicata. Os painéis sensoriais foram realizados em 1994, no Nabisco Food Center da Rutgers University (EUA), como parte de um projeto de pesquisa para o exército norte-americano. Os atributos avaliados estão agrupados por similaridade, conforme apresentado na Tabela 2.1.

Tabela 2.1: Atributos sensoriais

Identificação (ID)	Quesito	Atributo sensorial
1	Aparência	Proporção de molho na carne
2		Espessura visual do molho
3		Cor do molho (dados não avaliados)
4		Cor dos cubos da carne (dados não avaliados)
5		Uniformidade de tamanho e forma da carne
6	Sabor	Aroma de carne cozida
7		Aroma de caldo de carne
8		Aroma de carne crua
9		Aroma de proteína vegetal hidrolisada
10		Carne com sangue coagulado
11		Espessura
12		Aroma de carne queimada
13		Aroma de gordura
14	Sabores Básicos	Sal
15	Sensações	Sensação metálica
16		Sensação de calor
17	Textura	Viscosidade do molho
18		Elasticidade da carne
19		Densidade da carne
20		Coesividade inicial da carne
21		Firmeza da carne
22	Características de mastigação	Maciez da carne
23		Fibrosidade da carne
24		Estratificação da carne
25		Umidade da carne
26	Características Residuais	Película oleosa

PCA foi aplicada na matriz de correlação dos atributos padronizados. A matriz de correlação e a variância dos atributos são apresentadas no Apêndice. Os 24 atributos (os atributos 3 e 4 foram descartados da análise por apresentarem dados incompletos) foram tratados como variáveis e as amostras como observações (8 formulações × 8 julgadores × 4 repetições). Foram retidos 4 componentes principais, respondendo por 68% da variância. A inclusão de componentes adicionais não eleva a acurácia da classificação, de acordo com os testes realizados. Os pesos dos componentes foram manipulados através da equação (2), gerando os índices de importância *IA* para os 24 atributos, conforme apresentado na Tabela 2.2 (observe que os atributos são apresentados em ordem decrescente de importância).

Tabela 2.2: Pesos absolutos dos 4 componentes principais (CP) retidos e índice de importância (IA) em ordem decrescente de importância

ID do atributo	Pesos absolutos				*IA*
	CP1	CP2	CP3	CP4	
10	0,157	0,229	0,191	0,305	0,881
2	0,196	0,057	0,091	0,509	0,853
16	0,003	0,282	0,314	0,236	0,835
17	0,121	0,100	0,098	0,494	0,813
21	0,092	0,327	0,364	0,026	0,810
18	0,243	0,223	0,274	0,069	0,808
22	0,112	0,223	0,421	0,049	0,804
15	0,023	0,425	0,321	0,025	0,794
19	0,002	0,393	0,152	0,206	0,753
9	0,405	0,152	0,026	0,136	0,719
26	0,211	0,311	0,183	0,009	0,713
6	0,296	0,094	0,132	0,146	0,668
20	0,422	0,039	0,158	0,034	0,652
1	0,180	0,211	0,044	0,214	0,648
7	0,045	0,206	0,215	0,144	0,609
8	0,260	0,036	0,140	0,147	0,582
25	0,150	0,116	0,030	0,265	0,562
23	0,445	0,044	0,004	0,066	0,559
5	0,037	0,094	0,285	0,142	0,557
14	0,168	0,032	0,174	0,142	0,516
11	0,081	0,160	0,051	0,213	0,505
12	0,012	0,074	0,273	0,037	0,396
13	0,098	0,122	0,054	0,065	0,339
24	0,006	0,156	0,023	0,045	0,231

Na sequência, iniciou-se o processo iterativo de classificação e eliminação dos atributos (descrito na seção 2.4), gerando-se o gráfico de acurácia da Figura 2.2. A avaliação subjetiva deste gráfico mostra que o perfil de acurácia permanece estável à medida que os primeiros 11 atributos são removidos, e então existe uma queda considerável na acurácia

(ponto destacado no perfil e definido como ponto de corte). Tal ponto corresponde a 14 atributos retidos.

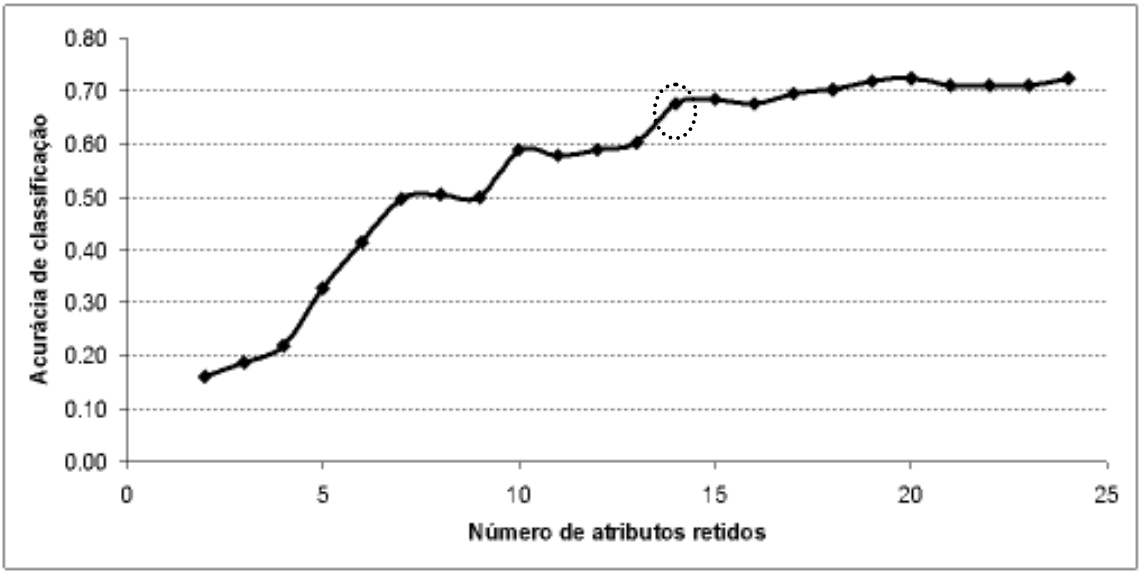


Figura 2.2: Perfil de acurácia gerado pela remoção de atributos

A Tabela 2.3 apresenta os valores de acurácia após cada eliminação de atributo (bem como a identificação do atributo eliminado). A primeira linha numérica desta tabela apresenta 0,723 como a acurácia de classificação obtida quando 24 atributos são considerados; na linha seguinte, o atributo 24 é eliminado, e a acurácia de classificação utilizando os 23 atributos remanescentes decresce para 0,711. O ponto de corte selecionado refere-se ao subconjunto formado por 14 atributos, os quais conduzem a uma acurácia de 0,676. Tal subconjunto (definido como subconjunto *SA*) é formado pelos seguintes atributos, em ordem decrescente de importância: 10, 2, 16, 17, 21, 18, 22, 15, 19, 9, 26, 6, 20 e 1. O subconjunto *AE* é constituído pelos 5 atributos removidos antes que o ponto de corte fosse atingido (atributos 7, 8, 25, 23 e 5), na Tabela 2.3.

Tabela 2.3: Identificação do atributo eliminado

Acurácia	Número de atributos retidos	ID do atributo eliminado
0,723	24	-
0,711	23	24
0,711	22	13
0,711	21	12
0,723	20	11
0,719	19	14
0,703	18	5
0,695	17	23
0,676	16	25
0,684	15	8
0,676	14	7
0,602	13	1
0,589	12	20
0,578	11	6
0,589	10	26
0,500	9	9
0,503	8	19
0,496	7	15
0,414	6	22
0,328	5	18
0,218	4	21
0,187	3	17
0,160	2	16
Atributos posicionados além do limite de eliminação		2 10

A enumeração das possíveis combinações incluiu os atributos *SA+AE* acima listados. O número de atributos a serem retidos variou de 14 a 19; o primeiro valor é arbitrado como o número mínimo de atributos a serem mantidos com base no ponto de corte, enquanto que 19 reflete a inclusão dos 5 atributos oriundos de *AE*. Por exemplo, a primeira enumeração testou a acurácia de todos os possíveis subconjuntos de 14 atributos formados a partir dos 19 atributos candidatos (*SA+AE*), resultando em um valor de 0,714. A Tabela 2.4 traz a maior acurácia gerada pela retenção de diferentes números de atributos. Este procedimento confirma que 16 atributos devem ser retidos, o que assegura uma acurácia de 0,723, valor idêntico ao obtido quando todos os atributos são utilizados. Os atributos selecionados são: 1, 2, 5, 6, 7, 8, 9, 10, 16, 17, 19, 20, 21, 23, 25 e 26. Observe que, dentre os atributos selecionados, figuram todos os atributos de aparência, bem como a maioria dos atributos de sabor.

Tabela 2.4: Variação da acurácia com o número de atributos retidos no procedimento de enumeração

Acurácia	Número de atributos retidos
0,714	14
0,719	15
0,723	**16**
0,723	17
0,719	18
0,719	19

Por fim, a acurácia do método proposto (0,723 com 16 atributos) foi comparada à máxima acurácia possível, obtida através da enumeração total de combinações geradas pelos 24 atributos (i.e., $\binom{24}{1}+...+\binom{24}{24}$). O máximo valor de acurácia possível é 0,742, obtido com 19 atributos (os 19 atributos responsáveis pela acurácia de 0,742 são diferentes dos 19 atributos selecionados pelo método, acima apresentados). Vale enfatizar que a enumeração total das possibilidades demandou 32 horas de processamento em PC 2.4 GHz, ao passo que a enumeração decorrente do método proposto demandou 10 minutos de análise.

2.6 CONCLUSÕES

Em painéis de avaliação sensorial, a existência de um número elevado de atributos conduz a procedimentos longos e caros, além de impor fadiga aos membros do painel. De tal forma, a seleção dos atributos mais relevantes constitui-se em uma importante etapa no planejamento desses painéis.

Este estudo apresentou um método para seleção de atributos na avaliação sensorial de alimentos. O método concilia duas metodologias de análise multivariada: análise de componentes principais (PCA) e análise discriminante (AD). Os pesos gerados pela PCA dão origem a um índice de importância dos atributos. A AD é então utilizada para classificação das amostras em diferentes classes, denotando formulações distintas. Um procedimento iterativo é desencadeado através da classificação das amostras e subsequente eliminação dos atributos menos relevantes, de acordo com a ordem definida pelos índices de importância. A acurácia de classificação é avaliada após cada eliminação de atributo, indicando um subconjunto de potenciais atributos a serem utilizados. Por fim, um procedimento de

enumeração identifica o melhor subconjunto de atributos a serem considerados em painéis sensoriais.

O método proposto permitiu reduzir o número de atributos em um painel de análise sensorial de cubos de carne ao molho de 24 para 16, mantendo-se os mesmos níveis de acurácia. Por fim, verificou-se que o resultado gerado pela metodologia proposta é similar à máxima acurácia possível, a qual é obtida de forma exaustiva por enumeração sobre a totalidade de atributos.

2.7 REFERÊNCIAS

ANZALDÚA-MORALES, A. A., **La evaluación sensorial de los alimentos en la teoria e la práctica**, Zaragoza: Editorial Acribia S.A., 1994.

CARBONELL, L.; IZQUIERDO, L.; CARBONELL, I. Sensory analysis of Spanish mandarin juices. Selection of attributes and panel performance, **Food Quality and Preference**, v.18, p.329–341, 2007.

CARBONELL, L.; IZQUIERDO, L.; CARBONELL, I.; COSTELL, E. (2008), Segmentation of food consumers according to their correlations with sensory attributes projected on preference spaces. **Food Quality and Preference**, v.19, p.71–78, 2008.

DIJKSTERHUIS, G.; FRØST, M. B.; BYRNE, D. V. Selection of a subset of variables: minimisation of Procrustes loss between a subset and the full set, **Food Quality and Preference**, v.13, p.89–97, 2002.

DUDA, R.; HART, P.; STORK, D. **Pattern Classification**, New York: Wiley-Interscience, 2nd ed, 2001.

ESTEBAN-DÍEZ, I.; GONZÁLEZ-SÁIZ, J.M.; PIZARRO, C. Prediction of sensory properties of espresso from roasted coffee samples by near-infrared spectroscopy, **Analytica Chimica Acta**, v.525, p.171–182, 2004.

FOGLIATTO, F.S.; ALBIN, S.L. A hierarchical method for evaluating products with quantitative and sensory characteristics, **IIE Transactions**, v.33, p.1081-1092, 2001.

GRANITTO, P.M.; BIASIOLI, F.; ENDRIZZI I.; GASPERI, F. Discriminant models based on sensory evaluations: Single assessors versus panel average, **Food Quality and Preference**, v.19, p.589-595, 2008.

GRANITTO, P.M.; GASPERI, F.; BIASIOLI, F.; TRAINOTTI, E.; FURLANELLO, C. Modern data mining tools in descriptive sensory analysis: A case study with a Random forest approach, **Food Quality and Preference**, v.18, p.681–689, 2007.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C., **Análise multivariada de dados**, New York: Bookman, 2006.

HEENAN, S. P.; DUFOUR, J. P.; HAMID, N.; HARVEY, W.; DELAHUNTY, C. M. Characterization of fresh bread flavor: Relationships between sensory characteristics and volatile composition. **Food Chemistry**, v.116, p.249–257, 2009.

JACKSON, J.E. **Principal Component and factor Analysis: Part I – Principal Components, Journal of Quality Technology**, v.12, p.201-213, 1980.

JACKSON, J.E. Principal Component and factor Analysis: Part II – Additional Topics Related to Principal Components. **Journal of Quality Technology**, v.13, p.46-58, 1981.

JOBSON, J. D. **Applied multivariate data analysis**. New York: Springer-Verlag, 1996.

JOHANSEN, S.M.B.; LAUGESEN, J.L.; JANHØJ, T.; IPSEN, R.H.; FRØST, M.B. Prediction of sensory properties of low-fat yoghurt and cream cheese from surface images, **Food Quality and Preference**, v.19, p.232–246, 2008.

KAROUI, R.; PILLONEL, L.; SCHALLER, E.; BOSSET, J.-O.; DE BAERDEMAEKER J. Prediction of sensory attributes of European Emmental cheese using near-infrared spectroscopy: A feasibility study. **Food Chemistry**, v.101, p.1121–1129, 2006.

KETTANEH, N.; BERGLUND, A.; WOLD, S. PCA and PLS in very large datasets. **Computational Statistics & Data Analysis**, v.48, p.69-85, 2005.

KRZANOWSKI, W. Selection of variables, and assessment of their performance, in mixed-variable discriminant analysis. **Computational Statistics and Data Analysis**, v.19, p.419-431, 1995.

KREUTZMANN, S.; SVENSSON, V.T.; THYBO, A.K.; BRO, R.; PETERSEN, M.A. Prediction of sensory quality in raw carrots (Daucus carota L.) using multi-block LS-ParPLS, **Food Quality and Preference**, v.19, p.609–617, 2008.

LEDAUPHIN, S.; HANAFI, M.; QANNARI, E.M. Assessment of the agreement among the subjects in fixed vocabulary profiling. **Food Quality and Preference**, v. 17, p. 277-280, 2006.

LEDAUPHIN, S.; POMMERET, D.; QANNARI, M. Application of hidden Markov model to products shelf lives. **Food Quality and Preference**, v.19, p.156–161, 2008.

MEILGAARD, M.C. ; CARR, B.T.; CIVILLE, G.V. **Sensory Evaluation Techniques**, 4ª ed, Boca Ratón, CRC Press, 1999.

MINGOTI, S. A., **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. UFMG, Belo Horizonte, 2005.

MURRAY, J. M.; DELAHUNTY, C. M.; BAXTER, I. A. Descriptive sensory analysis: past, present and future. **Food Research International**, v.34, p.461-471, 2001.

PERON, L. Statistical analysis of sensory profiling data: data reductionand generalised Procrustes analysis. **Food Quality and Preference**, v.11, p.155-157, 2000.

PIGGOTT, J. R.; SIMPSON, S. J.; WILLIAMS, S. A. R. Sensory analysis. International **Journal of Food Science and Technology**, v.33, p.7-18, 1998.

RENCHER, A. C. **Methods of Multivariate Analysis**, New York: Wiley, 1995.

SAHMER, K.; VIGNEAU, E.; QANNARI, E.M. A cluster approach to analyze preference data: choice of the number of clusters. **Food Quality and Preference**, v.17, p.257–265, 2006.

SAHMER, K.; QANNARI, E.M. Procedures for the selection of a subset of attributes in sensory profiling. **Food Quality and Preference**, v.19, p.141–145, 2008.

SEBER, G. A. F. **Multivariate observations**. New York: Wiley, 1984.

STONE, H.; SIDEL, J. **Sensory Evaluation Practices**, 2a Ed. San Diego: Academic Press, 1992.

STONE, H.; SIDEL, J.; OLIVER, S.; WOOLSEY, A.; SINGLETON, R.C. Sensory Evaluation by Quantitative Descriptive Analysis. **Food Technology**, v. 28, n.1, p. 24-34, 1974.

SUEYOSHI, T.; GOTO. M. Methodological comparison between DEA (data envelopment analysis) and DEA–DA (discriminant analysis) from the perspective of bankruptcy assessment. **European Journal of Operational Research**, v.199, n.2, p.561-575, 2009.

WESTADA, F.; HERSLETHA, M.; LEAA, P.; MARTENS, H. Variable selection in PCA in sensory descriptive and consumer data. **Food Quality and Preference**, v.14, p.463–472, 2003.

ZANINE, A. M.; DIAS, P. F.; SOUTO, S. M.; FERREIRA, T, D. J.; SANTOS, E. M.; PINTO, L. F. B. Evaluation of the grass-tanzania ("Panicum maximum") using multivariate analyses. **Revista Brasileira de Saúde e Produção Animal**, v.9, p.179-189, 2008.

	1	2	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	Variância
1	1,00	0,08	0,02	0,24	-0,01	0,21	0,02	0,09	-0,04	-0,03	-0,33	0,14	0,25	0,21	0,18	0,11	0,12	0,25	0,00	0,03	0,23	0,22	0,00	0,15	**1,670**
2	-0,08	1,00	0,09	0,08	-0,04	-0,25	0,04	0,28	-0,05	0,10	0,13	0,03	0,00	-0,11	0,57	-0,23	0,10	0,26	-0,15	-0,05	0,27	0,23	-0,09	0,06	**1,604**
5	0,02	0,09	1,00	0,16	-0,17	0,01	-0,04	-0,02	-0,22	-0,12	0,07	-0,03	-0,04	-0,17	0,03	0,16	0,09	0,17	0,23	0,22	0,03	-0,06	0,05	0,00	**1,985**
6	0,24	0,08	0,16	1,00	0,02	-0,32	0,04	0,08	-0,05	0,09	0,32	-0,30	0,01	-0,04	-0,02	0,01	0,03	0,44	-0,04	0,25	0,42	0,42	0,21	0,12	**1,036**
7	0,01	-0,04	-0,17	0,02	1,00	-0,09	0,03	-0,07	0,00	0,11	-0,08	-0,10	-0,13	0,04	0,00	-0,13	-0,19	0,07	0,01	-0,07	0,15	0,25	0,08	0,11	**0,504**
8	-0,21	-0,25	0,01	-0,32	-0,09	1,00	-0,02	-0,12	0,21	0,15	-0,11	0,22	0,11	0,14	-0,18	0,22	-0,10	-0,35	0,02	-0,16	-0,33	-0,30	0,04	-0,11	**0,605**
9	0,02	0,04	-0,04	0,04	0,03	-0,02	1,00	-0,17	0,04	0,07	-0,07	0,03	-0,26	-0,03	0,07	-0,04	-0,10	0,04	-0,08	0,00	0,01	-0,01	-0,01	0,03	**0,455**
10	0,09	0,28	-0,02	0,08	-0,07	-0,12	-0,17	1,00	0,03	0,14	0,33	-0,06	0,41	0,08	0,16	-0,20	0,15	0,20	-0,02	-0,01	0,23	0,11	-0,22	-0,02	**0,594**
11	-0,04	-0,05	-0,22	-0,05	0,00	0,21	0,04	0,03	1,00	0,18	0,09	-0,07	0,25	0,37	-0,04	0,01	0,12	-0,05	0,01	0,03	-0,05	-0,09	-0,05	0,05	**0,331**
12	-0,03	0,10	-0,12	0,09	0,11	0,15	0,07	0,14	0,18	1,00	0,14	0,00	0,22	0,26	0,08	-0,02	-0,12	-0,08	-0,07	-0,09	0,02	0,02	-0,01	-0,10	**0,801**
13	0,33	0,13	0,07	0,32	-0,08	-0,11	-0,07	0,33	0,09	0,14	1,00	-0,08	0,46	0,20	0,10	-0,08	0,14	0,26	-0,07	0,03	0,25	0,12	0,16	0,28	**0,559**
14	-0,14	0,03	-0,03	-0,30	-0,10	0,22	0,03	-0,06	-0,07	0,00	-0,08	1,00	0,12	0,11	0,02	0,04	0,03	-0,27	-0,08	-0,25	-0,17	-0,19	-0,06	-0,07	**1,446**
15	0,25	0,00	-0,04	0,01	-0,13	0,11	-0,26	0,41	0,25	0,22	0,46	0,12	1,00	0,43	0,00	0,01	0,26	-0,06	0,04	-0,10	0,00	-0,11	0,04	0,15	**0,948**
16	0,21	-0,11	-0,17	-0,04	0,04	0,14	-0,03	0,08	0,37	0,26	0,20	0,11	0,43	1,00	-0,17	-0,05	0,13	-0,02	0,03	-0,12	0,02	0,01	0,00	0,14	**1,130**
17	-0,18	0,57	0,03	-0,02	0,00	-0,18	0,07	0,16	-0,04	0,08	0,10	0,02	0,00	-0,17	1,00	-0,22	0,07	0,12	-0,11	-0,10	0,13	0,06	0,05	0,08	**1,937**
18	-0,11	-0,23	0,16	0,01	-0,13	0,22	-0,04	-0,20	0,01	-0,02	-0,08	0,04	0,01	-0,05	-0,22	1,00	0,20	-0,26	0,48	0,22	-0,43	-0,46	-0,08	-0,06	**2,223**
19	0,12	0,10	0,09	0,03	-0,19	-0,10	-0,10	0,15	0,12	-0,12	0,14	0,03	0,26	0,13	0,07	0,20	1,00	0,05	0,44	0,33	-0,05	-0,12	-0,16	0,00	**2,069**
20	0,25	0,26	0,17	0,44	0,07	-0,35	0,04	0,20	-0,05	-0,08	0,26	-0,27	-0,06	-0,02	0,12	-0,26	0,05	1,00	0,06	0,41	0,86	0,67	0,21	0,10	**10,687**
21	0,00	-0,15	0,23	-0,04	0,01	0,02	-0,08	-0,02	0,01	-0,07	-0,07	-0,08	0,04	0,03	-0,11	0,48	0,44	0,06	1,00	0,49	-0,13	-0,23	-0,26	0,02	**0,762**
22	0,03	-0,05	0,22	0,25	-0,07	-0,16	0,00	-0,01	0,03	-0,09	0,03	-0,25	-0,10	-0,12	-0,10	0,22	0,33	0,41	0,49	1,00	0,20	0,05	0,03	0,05	**3,553**
23	0,23	0,27	0,03	0,42	0,15	-0,33	0,01	0,23	-0,05	0,02	0,25	-0,17	0,00	0,02	0,13	-0,43	-0,05	0,86	-0,13	0,20	1,00	0,83	0,29	0,09	**10,448**
24	0,22	0,23	-0,06	0,42	0,25	-0,30	-0,01	0,11	-0,09	0,02	0,12	-0,19	-0,11	0,01	0,06	-0,46	-0,12	0,67	-0,23	0,05	0,83	1,00	0,31	0,00	**12,242**
25	0,00	-0,09	0,05	0,21	0,08	0,04	-0,01	-0,22	-0,05	-0,01	0,16	-0,06	0,04	0,00	0,05	-0,08	-0,16	0,21	-0,26	0,03	0,29	0,31	1,00	0,07	**1,571**
26	0,15	0,06	0,00	0,12	0,11	-0,11	0,03	-0,02	0,05	-0,10	0,28	-0,07	0,15	0,14	0,08	-0,06	0,00	0,10	0,02	0,05	0,09	0,00	0,07	1,00	**0,491**

3 ARTIGO 2 – Método Baseado na Mineração de Dados para Identificação de Atributos Discriminantes em Perfis Sensoriais

Karina Rossini

Michel J. Anzanello

Flávio S. Fogliatto

Resumo

A seleção dos atributos de um conjunto de atributos candidatos, a serem avaliados em uma análise sensorial, é uma importante etapa no planejamento de painéis sensoriais. Na seleção, deseja-se reduzir a lista de atributos a serem avaliados, evitando, assim, a imposição de fadiga aos membros do painel, minimizando custos e tempo. Em algumas aplicações, o objetivo é manter os atributos que são relevantes e não redundantes na caracterização sensorial dos produtos. Neste artigo, no entanto, o interesse é manter os atributos que melhor discriminam os produtos. Para tanto, é apresentado um método de mineração de dados para seleção de atributos descritivos em painéis sensoriais, tais em Análise Descritiva Quantitativa. O método proposto é implementado utilizando Análise de Componentes Principais e a técnica de classificação dos *k* vizinhos mais próximos, em conjunto com a análise do Pareto Ótimo. Os objetivos são (*i*) identificar o conjunto de atributos que melhor discrimina as amostras analisadas pelo painel, e (*ii*) indicar o grupo de julgadores que fornecem avaliações consistentes. O método é ilustrado através de um estudo de caso onde cubos de carne moída bovina, usada como ração de combate pelo Exército Americano, são caracterizados por painéis sensoriais utilizando o protocolo *Spectrum*.

Palavras-chave: Seleção de atributos, Atributos discriminantes, Classificação de amostras, Ferramentas de mineração de dados

Abstract

Selection of attributes from a group of candidates to be assessed through sensory analysis is an important issue when planning sensory panels. In attribute selection it is desirable to reduce the list of those to be presented to panelists to avoid fatigue, minimize costs and save time. In some applications the goal is to keep attributes that are relevant and non-redundant in the sensory characterization of products. In this paper, however, we are

interested in keeping attributes that best discriminate between products. For that we present a data mining-based method for attribute selection in descriptive sensory panels, such as those used in the Quantitative Descriptive Analysis. The proposed method is implemented using Principal Component Analysis and the k-Nearest Neighbor classification technique, in conjunction with Pareto Optimal analysis. Objectives are (i) to identity the set of attributes that best discriminate samples analyzed in the panel, and (ii) to indicate the group of panelists that provide consistent evaluations. The method is illustrated through a case study where beef cubes in stew, used as combat ration by the American Army, are characterized in sensory panels using the Spectrum protocol.

Keywords: Attribute selection, Discriminant attributes, Sample classification, Data mining tools

3.1 INTRODUÇÃO

Os métodos de Análise Descritiva (AD) têm como objetivo fornecer o perfil sensorial dos produtos. Em essência, os protocolos de AD para a avaliação de amostras utilizam um conjunto de atributos tipicamente grande. As amostras são avaliadas individualmente, e os resultados são expressos em escala numérica contínua. Há um único valor indicando a intensidade do atributo para cada amostra assim, dados do painel sensorial podem ser tratados como dados quantitativos.

Embora amplamente utilizados (Murray *et al.*, 2001), os métodos de AD apresentam algumas limitações. Primeiro, o número de atributos a ser avaliado pelos painlistas é extenso; alguns protocolos de AD podem apresentar até 30 atributos (Carbonell *et al.*, 2007). Como resultado, a coleta de dados tende a ser cansativa, demorada e onerosa. Em segundo lugar, não é possível afirmar que o perfil completo das amostras irá incluir os atributos através dos quais as amostras possam ser discriminadas, embora este seja o principal objetivo da coleta de dados (Granitto *et al.*, 2007). Terceiro, os métodos AD não oferecem qualquer forma estruturada de pontuação para os julgadores.

A seleção de atributos é um importante tema de pesquisa na área de avaliação sensorial. A seleção pode destinar-se a (*i*) identificar um subconjunto de atributos não redundantes que melhor descrevem os produtos, ou (*ii*) encontrar atributos que melhor discriminam os produtos.

Com relação ao objetivo (*i*), autores como Dijksterhuis *et al.*, (2002), Westad *et al.*, (2003) e Sahmer e Qannari (2008) propuseram o uso de vários métodos de projeção

multivariada, para selecionar um subconjunto de atributos relevantes e/ou não redundantes de um grupo maior. No que diz respeito ao objetivo (*ii*), e metodologicamente alinhado com as proposições do presente artigo, Granitto *et al.* (2007) introduziram o uso de ferramentas de mineração dados (mais especificamente *Random forests heuristics* ou heurísticas de florestas aleatórias) para selecionar os atributos que melhor discriminam os produtos. Em todos os casos, perfis sensoriais foram usados para caracterizar os produtos e as limitações listadas acima foram parcialmente resolvidas.

Neste trabalho, propõe-se o uso combinado de métodos de projeção multivariada e ferramentas de mineração de dados para selecionar atributos relevantes no perfil sensorial. Atributos relevantes são aqueles que melhor discriminam os produtos avaliados em um painel. O método sugerido é implementado em seis etapas. Primeiro, os participantes são classificados utilizando um índice de consistência; aqueles cujas avaliações diferem dos demais integrantes do grupo têm seus dados omitidos do conjunto de dados. Este procedimento é realizado utilizando uma abordagem *''deixe um julgador de fora por vez''*. *S* conjuntos de dados são produzidos, cada um composto de avaliações a partir de um subgrupo de julgadores; o número mínimo de participantes em um subgrupo é definido pelo usuário.

Etapas remanescentes são implementadas *S* vezes, uma para cada conjunto de dados obtidos no passo inicial. Para cada conjunto de dados, a Análise de Componentes Principais (PCA) é aplicada e os índices de importância de atributos com base nos pesos da PCA são calculados. Posteriormente, classificam-se os produtos no conjunto de dados usando o algoritmo *k* Vizinhos mais Próximos (KNN ou *k-Nearest Neighbor*) e calcula-se a acurácia da classificação. O atributo com o menor índice de importância é removido do conjunto de dados, os produtos são novamente classificados e um novo valor de acurácia é produzido. O processo é repetido até que haja apenas um atributo restante. Os resultados da acurácia são plotados em um gráfico e todo o procedimento é repetido para o próximo conjunto de dados. Os resultados de cada iteração são reunidos em um único gráfico de acurácia, e o melhor conjunto de atributos e julgadores é determinado utilizando a Análise de Pareto Ótimo.

Destacam-se três contribuições relevantes no método proposto. Primeiro, o método de seleção de atributos identifica, simultaneamente, os atributos discriminantes e os julgadores consistentes. Mais especificamente, o poder discriminatório dos diversos subgrupos de atributos é medido em função das avaliações realizadas pelos diferentes grupos de julgadores. O objetivo é identificar o melhor conjunto de atributos discriminantes,

mantendo os julgadores cujas avaliações melhoram a capacidade de discriminação dos produtos pelos atributos. Outras proposições de seleção de atributos encontradas na literatura concentram-se exclusivamente na minimização dos atributos retidos, como em Granitto *et al.* (2007).

Segundo, a Análise de Componentes Principais é combinada com o algoritmo de classificação KNN para obter um método eficiente de seleção de atributos. Nosso método utiliza a técnica de classificação KNN devido ao seu bom desempenho em aplicações práticas no campo de mineração de dados, simplicidade conceitual e ampla disponibilidade em pacotes computacionais; ver Chaovalitwongse *et al.* (2007) e Anzanello *et al.* (2009).

Terceiro, utiliza-se a análise de Pareto Ótimo (PO) para identificar um número limitado de soluções distintas que maximiza a acurácia de classificação e minimiza tanto o número de atributos retidos quanto de julgadores. PO tem sido empregado em uma grande variedade de aplicações, tais como a análise dos ciclos de vida de produtos químicos em Azapagic (1999), *scheduling* de operações de manufatura em Taboada e Coit (2008).

Para ilustrar o método proposto de seleção de atributos, este é aplicado em um estudo de caso envolvendo cubos de carne ao molho, embalados em *pouches* termoestáveis, para uso militar (Fogliatto *et al*, 1999). O produto foi produzido em uma planta piloto, localizada na Rutgers University, EUA. Oito produtos diferentes foram avaliados em relação a 26 atributos sensoriais por um painel sensorial com nove membros. O método apresentado reduziu o número de julgadores e atributos necessários para discriminar os produtos para 5 e 17, respectivamente.

O restante deste trabalho está organizado da seguinte forma. A Seção 3.2 apresenta uma breve revisão da literatura sobre propostas de seleção de atributos das abordagens no campo sensorial, e introduz algumas das ferramentas aplicadas no método proposto. A Seção 3.3 apresenta a seqüência de passos para aplicação do método. A Seção 3.4 mostra os resultados numéricos da aplicação do método no estudo de caso, seguido de conclusões na Seção 3.5.

3.2 REFERENCIAL TEÓRICO

A seleção de variáveis e características é um importante tema de pesquisa em áreas onde frequentemente grandes conjuntos de dados estão envolvidos, tais como Genética e

Linguística. O objetivo de selecionar variáveis é melhorar o desempenho dos preditores, e que melhor descrever o processo de geração de dados. Estratégias para a seleção de variáveis têm sido revisadas por Guyon e Elisseeff (2003), com ênfase especial em métodos de classificação. Tais métodos são baseados na ordenação de variáveis de acordo com um indice de importância, descartando aquelas com o menor índice de pontuação e reduzindo, assim, o conjunto de dados (Gauchi & Chagnon, 2001).

Em análise sensorial, a classificação é comumente realizada através de ANOVA usando os *F*-valores dos atributos como índice de importância. Se um índice multivariado é desejado, métodos de projeção como a Análise de Componentes Principais (PCA) ou STATIS podem ser utilizados. Alguns trabalhos referentes à seleção de atributos para fins de discriminação em análise sensorial são apresentados aqui. Com exceção de Granitto *et al.* (2007) onde as ferramentas de mineração de dados são aplicadas, todas as abordagens são baseadas em combinações de técnicas de análise multivariada e análise de variância.

A Análise Discriminante Linear (ADL) considera-se que os dados seguem uma distribuição normal multivariada, com matriz de covariâncias comum para todas as categorias (Ripley, 1996). A distância de Mahalanobis de cada objeto a partir de centróides das categorias é computada, e os objetos são atribuídos à categoria com menor distância. O delimitador entre as duas categorias é uma função linear, que no caso de duas variáveis independentes, pode ser uma linha reta. A classificação dos atributos na ADL é, geralmente, realizada utilizando a taxa de erro de previsão em uma amostra de validação como índice de importância. Uma aplicação tradicional de ADL na seleção de atributos pode ser encontrada em Rason *et al.,* (2007), enquanto Granitto *et al.,* (2008) estendem o uso da ADL com métodos mais elaborados. Em comparação com o método aqui proposto, procedimentos de seleção atributos baseados em ADL não testam diferentes subconjuntos de atributos discriminantes em busca do melhor conjunto. O procedimento identifica os atributos com base no desempenho de uma amostra de validação e escolhe um ponto de corte para determinar o subconjunto de atributos que devem ser conservados. Por outro lado, de forma iterativa, avalia a qualidade da classificação, por testar extensivamente diferentes subconjuntos de atributos, em busca da melhor acurácia de classificação.

A ANOVA pode ser utilizada para determinar se existem diferenças significativas entre os produtos ou julgadores. No entanto, a aplicação de ANOVA para cada atributo sensorial aumenta a probabilidade de erro Tipo I (Tabachnick & Fidell, 1996). Para superar

isso, a ANOVA pode ser usada em conjunto com PCA para fins de seleção de atributos. Uma abordagem usual consiste na aplicação da ANOVA sobre escores obtidos a partir da PCA em dados originais da matriz sensorial, como reportado em Chabanet (2000) Westad *et al.* (2003). Neste caso, a matriz de dados apresenta combinações de produtos e julgadores nas linhas, e atribuem as avaliações nas colunas. Em casos onde efeitos significativos estão presentes, os testes de Fisher LSD *post hoc* podem ser empregados para explorar as diferenças entre os produtos individuais ou julgadores. Luciano e Naes (2009) referem-se a esta abordagem como PCA-ANOVA.

Alternativamente, a ANOVA pode ser utilizada em cada atributo da matriz de dados separadamente, seguido da PCA nas matrizes de efeitos principais e interações. Esse procedimento, conhecido como ASCA (Jansen *et al.*, 2006), usa PCA para interpretar os resultados da ANOVA. O método ASCA e o da PCA seguida pela ANOVA são comparados por Luciano e Naes (2009) em um conjunto de dados de análise sensorial de um doce, com resultados semelhantes. Em oposição ao método proposto, nenhuma das abordagens acima testa o desempenho dos subconjuntos de atributos quanto ao poder de classifição, nem verificam o melhor subconjunto de participantes.

Granitto *et al.* (2007) propôs a utilização das *random forests* (RFs) para selecionar atributos discriminantes em amostras de queijo. RFs são conjuntos heurísticos de árvores de decisão criadas de tal forma que as diferenças entre as árvores sejam maximizadas. A função discriminante derivada de RFs atribui pesos de importância aos atributos; esses pesos podem ser usados para selecionar um número reduzido de atributos, porém, mantendo o poder discriminatório da função em um nível desejado. Os autores foram os primeiros a propor o uso de ferramentas de mineração de dados no contexto da caracterização sensorial, no entanto, em sua proposta, a seleção de julgadores não é avaliada simultaneamente à seleção de atributos, como proposto neste artigo. Além disso, a técnica de RFs é uma ferramenta de classificação bastante complexa se comparada com o algoritmo KNN aqui proposto e não está disponível na maioria dos pacotes estatísticos.

Uma ampla pesquisa teórica de seleção de variáveis é relatada em Liu e Yu (2005), enquanto a comparação dos algorítmos para esse fim é apresentado por Kudo e Sklansky (2000): o melhor método combina a regra *leave-one-out* e a técnica de classificação KNN, corroborando com nossa proposta.

Algumas abordagens clássicas de seleção de variáveis são decorrentes da indústria alimentícia, embora não diretamente relacionadas com seleção de atributos sensoriais. Muitas delas utilizam PCA para identificar as variáveis com maior variância e, em seguida, aplicam ferramentas de classificação e *cluster* (por exemplo, técnicas de mineração de dados, discriminantes e *cluster*), utilizando as variáveis selecionadas; ver Mallet *et al.,* (1998) e Guo *et al.,* (2002). Um método para reduzir a dimensionalidade na fabricação industrial de vinho foi relatado por Urtubia *et al.,* (2007): PCA foi inicialmente utilizada para selecionar as variáveis com a maior informação sobre as interaçãos metabólitas, e classes com comportamentos semelhantes foram, então, geradas pela aplicação da técnica de *cluster* (*K-means*) no conjunto de dados de menor dimensão. Em Camara *et al.,* (2006), PCA foi associado com análise discriminante para diferenciar e classificar vinhos, os coeficientes das funções discriminantes foram usados para identificar as variáveis relevantes.

Um estudo semelhante foi realizado por Rebolo *et al.,* (2000) para testar a autenticidade dos vinhos produzidos em uma região específica. Variáveis que descrevem as substâncias químicas foram inicialmente selecionadas através das técnicas de análise de *cluster* e PCA seguidas por uma análise bayesiana passo a passo. O mesmo estudo também visava classificar os vinhos em classes de acordo com sua origem. Com finalidades idênticas, Marini *et al.,* (2006) estudaram duas técnicas de classificação, denominadas *Soft Independent Modeling of Class Analogies* (SIMCA) e *Unequal Class Modeling* (UNEQ) para verificar a autenticidade de vinhos italianos. SIMCA descreve as semelhanças dos produtos em uma categoria por meio da PCA (Wold & SjoStrom, 1977), enquanto o UNEQ é um modelo de classe normal multivariada assumindo uma dispersão individual (ou seja, dispersão desigual) de cada classe, similar a uma Análise Discriminante Quadrática (Derde & Massarte, 1986). A identificação das variáveis relevantes foi feita por meio de uma Análise Discriminante Linear *Stepwise*. Em Capron *et al.,* (2007), um conjunto de dados de vinhos, com 63 variáveis de processo, foi avaliado usando árvores de decisão e modificações na regressão de mínimos quadrados parciais (*Partial Least Square* - PLS) para identificar as variáveis mais importantes destinadas a classificação dos vinhos em 4 classes diferentes (países de origem).

Na sequência, algumas informações sobre duas das ferramentas analíticas utilizadas em nosso método, o algorítmo de classificação KNN e a análise de Pareto Ótimo, são apresentadas.

O KNN é uma técnica de mineração de dados para classificar objetos. A determinação da classe a qual um objeto pertence está baseada nos seus vizinhos mais próximos do espaço de variáveis. KNN é um dos algoritmos mais simples para classificação de observações (Duda *et al.,* 2001).

Considere as observações em um conjunto *J*-dimensional, correspondendo aos *J* atributos, e duas classes de produtos (A ou B). O objetivo é classificar uma nova observação em A ou B com base apenas em atributos. Consideram-se os *k*-vizinhos mais próximos da nova observação, a proximidade é medida através da distância Euclidiana. Para cada um dos *k*-vizinhos identificar a classe A ou B. Uma forma de classificar a nova observação é por maioria: a observação nova está na classe A, se a maioria dos seus *k*-vizinhos mais próximos está na classe A. Se $k = 1$, então a observação é simplesmente atribuída à classe do seu vizinho mais próximo. O número de vizinhos *k* (*k* é um número positivo, tipicamente pequeno), é dado pela maximização da acurácia de classificação no conjunto de dados onde a classe de cada observação é conhecida. Mais detalhes sobre a técnica de classificação KNN pode ser encontrado em Wu *et al.* (2008).

KNN apresenta vantagens de ser conceitualmente mais simples e mais intuitivo do que outras técnicas de classificação, além de amplamente disponível em pacotes de software. Além disso, KNN exige apenas um parâmetro, *k*, e a acurácia da classificação não é muito sensível quanto à escolha deste dentro de um intervalo razoável. Devido à sua simplicidade, KKN tem sido aplicado em uma ampla variedade de contextos, incluindo o reconhecimento paterno em Weiss *et al.* (1999), detecção de atividade cerebral anormal em Chaovalitwongse *et al.* (2007) e a classificação de lotes de produção em Anzanello *et al.* (2009).

A análise de Pareto Ótimo (PO) identifica um conjunto distinto de soluções em aplicações com múltiplas funções objetivo. Estas funções frequentemente não apresentam uma solução única, mas um conjunto de soluções adequadas. Aplicações em seleção de variáveis onde a medida de desempenho da classificação é maximizada e o número de variáveis retidas é minimizado são exemplos de cenários com várias soluções possíveis.

Devido a sua aplicabilidade prática, a análise de PO tem sido amplamente integrada a algoritmos e propostas para fins de otimização, como relatado em Deb *et al.,* (2002), e Deb *et al.,* (2002A). Soluções identificadas pelo Ótimo de Pareto são nomeadas não-dominadas, o que significa que não podem ser ultrapassadas por outras soluções vizinhas nos objetivos

avaliados. Isso resulta uma redução significativa do número de soluções possíveis, o que faz com que a análise possa ser focada em um pequeno conjunto de soluções efetivamente melhor. Estas soluções são tipicamente ilustradas em um limite chamado fronteira de Pareto. A identificação da melhor solução pode depender de informação subjetiva e pode tornar-se complexa à medida que o número de funções objetivo aumenta; ver Horn *et al.,* (1994) Zitzler e Thiele (1999), e Taboada e Coit (2008) para mais detalhes.

3.3 MÉTODO PROPOSTO

O método para selecionar os atributos que melhor discriminam produtos baseia-se em seis etapas operacionais: 1. Medir a consistência dos julgadores utilizando um índice apropriado. 2. Aplicar uma técnica multivariada no conjunto de dados consistindo de atributos sensoriais. Em nosso método, a Análise de Componentes principal (PCA) é utilizada, mas outras técnicas podem ser consideradas. 3. Computar um vetor de índice importância dos atributos com base nos pesos da PCA. 4. Classificar o conjunto de dados sensoriais através da técnica dos *k*-vizinhos mais próximos (KNN) e, calcular a acurácia de classificação. Em seguida, eliminar o atributo com o menor índice de importância, classificar o conjunto de dados novamente, e recalcular a acurácia. Continuar esse processo iterativo até que haja apenas um atributo restante. 5. Construir um gráfico de acurácia. 6. Remover o julgador menos consistente e realizar uma nova seleção de atributos, repetir os passos 2-6. Estas medidas operacionais estão detalhadas nas seções a seguir; os códigos do Matlab utilizados para realizar as análises são apresentados no apêndice.

Considere que p (p=1,...,P) denote os julgadores, j (j=1,...,J) os atributos, e i (i =1,...,I) os produtos analisados em um painel sensorial. As avaliações dos membros do painel, para um produto em relação a todos os atributos, podem ser repetidas d (d=1,...,D) vezes. Considere uma matriz $\mathbf{X}$ de dados sensoriais composta de (P x I x D) linhas e J colunas, com elemento $x_{pid,j}$. A matriz $\mathbf{X}$ pode ser decomposta em P matrizes individuais $\mathbf{X}_p$, cada uma contendo avaliações do julgador P para os J atributos e produtos com D repetições. O objetivo do método é o de classificar corretamente as observações (P x I x D) de $\mathbf{X}$ em I classes de produtos pela escolha adequada dos atributos e julgadores consistentes.

Passo 1: Medida de consistência dos julgadores

Medir a consistência dos julgadores usando um índice apropriado. Nós sugerimos utilizar o índice associado à média ponderada da configuração proposto por Ledauphin *et al.,* (2006). Este índice tem várias características interessantes. Primeiro, efeitos de localização e de disperssão nas avaliações dos julgadores são removidos através de pré-tratamento da matriz de dados. Segundo, com um quadro analítico semelhante ao método STATIS (Lavit *et al.,* 1994), o índice evidencia julgadores que apresentam avaliações dos atributos diferentes dos demais integrantes do painel sensorial. Em terceiro lugar, a aplicação do índice requer cálculos simples, tal como apresentado a seguir.

A centralização e redução da matriz de dados $\mathbf{X}_p$ é o primeiro passo. Para tanto, subtrair todas as entradas da coluna pela média da coluna para obter uma matriz de dados centrada $\mathbf{Xc}_p$. Em seguida, obter uma matriz $\mathbf{Y}_p$ multiplicando cada matriz $\mathbf{Xc}_p$ por um escalar $\theta_p = 1/\sqrt{t_p}$, onde t_p é a soma do quadrado de todas as entradas no $\mathbf{Xc}_p$ ou seja, $t_p = traço\left(\boldsymbol{Xc}_p^t\right)\boldsymbol{Xc}_p$ com $\mathbf{Xc}_p{}^t$ denotando a transposição de $\mathbf{Xc}_p$. Formalmente, $\boldsymbol{Y_p} = \theta_p \boldsymbol{Xc}_p$.

A configuração da média ponderada do conjunto de dados sensoriais é obtida da seguinte forma. Considere as matrizes $\mathbf{Y}_k$ e $\mathbf{Y}_l$ dos julgadores k e l. Determine uma matriz $\boldsymbol{S}$ (P x P) com entradas correspondentes a uma medida de similaridade entre os julgadores K e l dada por $S_{kl} = (1 + t_{kl})/2$, onde $t_{kl} = traço(\boldsymbol{Y}_k^l \boldsymbol{Y}_l)$, para k , l=1,. . . , P. Determine o autovetor correspondente ao maior autovalor de **S**, ou seja, $\beta^t = \left[\beta_1, \ldots, \beta_p\right]$, sendo $\sum_{p=l}^{p} \beta_p = 1$.

A configuração da média ponderada é uma matriz compromisso **C**, que leva em conta o desempenho dos julgadores. Formalmente, $\boldsymbol{C} = \sum_{p-l}^{p} \beta_p \boldsymbol{Y}_p$. O índice de desempenho para o julgador p é dado por:

$$\alpha_p = \frac{traço(\boldsymbol{Y}_p^t \boldsymbol{C})}{\sqrt{traço(\boldsymbol{C}^t \boldsymbol{C})}} \qquad (1)$$

Os valores alfa variam no intervalo de -1 a +1. Um julgador com um α próximo a -1 está em completo desacordo com o restante do painel, enquanto um membro do painel, com um valor α igual a 1 está em perfeito acordo com o resto do grupo. Os julgadores são, então, classificados de acordo com os valores de alfa, e o julgador com o menor α_p é o primeiro a ser retirado da análise. Cada remoção de julgador é seguida de um processo de seleção de atributos, detalhado nas etapas 2-6.

Passo 2: Aplicar uma técnica multivariada no conjunto de dados consistindo os atributos sensoriais

Caracterizar a relação entre os atributos da matriz **X** usando uma técnica multivariada, considerando atributos como variáveis na análise. Recomendamos o uso de PCA na matriz **X**. As saídas de interesse do PCA são os componentes pesos w_{jr} e o percentual da variância explicada por cada componente r retido (r=1,...,R). O número de componentes retidos R é definido com base na quantidade de variância explicada por eles, como em Montgomery *et al.* (2001).

Etapa 3: Gerar índices de importância para os atributos (z)

Gerar um índice de importância para os atributos. Este índice é utilizado para orientar a remoção de atributos não relevantes para fins de classificação, conforme proposto em Anzanello *et al.* (2009). O índice do atributo j é denotado por z_j, j=1,...,J. Quanto maior o valor de z_j, mais importante é este atributo para classificar as observações em classes de produtos.

O índice z_j é gerado com base nos pesos w_{jr} da PCA como na equação (2). Atributos com pesos grandes são preferidos pois, de acordo com Duda *et al.* (2001), tais atributos devem conduzir a uma melhor discriminação das observações em classes de produtos, embora exceções podem ocorrer.

$$Z_j = \sum_{r=1}^{R} |w_{jr}|, \quad j = 1, \dots, J \qquad (2)$$

Passo 4: Classificar o conjunto de dados usando KNN e eliminar os atributos irrelevantes e ruidosos

Categorizar o conjunto de dados das observações *(P* x *I* x *D)* em *I* classes de produtos com todos os *J* atributos usando KNN e calcular a acurácia de classificação. A acurácia é definida como a razão entre o número de classificações corretas e o número de classificações realizadas. O parâmetro k para o algoritmo KNN é selecionado através da validação cruzada sobre o conjunto de dados sensoriais, como no Chaovalitwongse *et al.* (2007).

A eliminação do atributo começa por identificar o atributo com o menor z_j. Remover o atributo selecionado, executar uma nova classificação utilizando o KNN nos *J*-1 atributos

restantes e computar a acurácia da classificação. Este procedimento é repetido removendo o próximo atributo com o menor z_j e aplicando KNN sobre os demais atributos, até que haja apenas um atributo remanescente.

Passo 5: Construa um gráfico de acurácia

Construir um gráfico relacionando a acurácia da classificação com o número de atributos retidos. Neste caso, a acurácia de classificação é o unico critério de otimização considerado, a máxima acurácia indica o melhor subconjunto de atributos que devem ser conservados para a classificação. No caso de ter subconjuntos alternativos com valores de acurácia idêntica, escolher aquele com o menor número de atributos retidos.

Passo 6: Remover o julgador menos consistente e executar uma nova seleção de atributos

Remover o julgador com o menor α_p e repetir os passos 2 a 6 para os dados sensoriais compostos dos J atributos originais e julgadores restantes. Adicionar o novo perfil de acurácia ao gráfico de acurácia da etapa 5. Note que, a eliminação de cada julgador, leva a seleção de um novo atributo com base nas avaliações dos julgadores restantes, uma das saídas do procedimento será um conjunto de perfis de acurácia. Repetir este procedimento iterativo até que um limite inferior de julgadores restante é alcançado.

A solução final é obtida através da identificação da acurácia (pico) máxima global no conjunto de perfis de acurácia, como exemplificado na Fig. 3.1. O pico identifica o melhor grupo de julgadores e o melhor subconjunto de atributos a serem considerados nos procedimentos de classificação. A análise Pareto Ótimo (PO) também pode ser usada para identificar um número limitado de soluções que maximizam a acurácia de classificação e minimizam o número de atributos retidos. Tais análises podem ser particularmente úteis quando o gráfico apresenta multiplos picos.

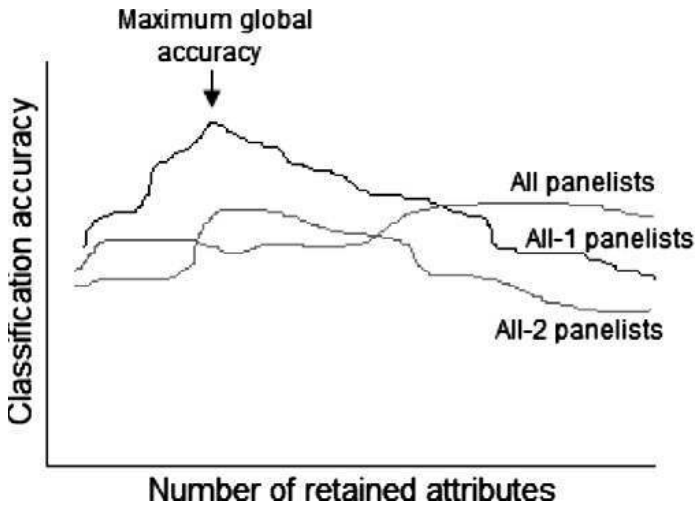


Figura 3.1: Perfil de acurácia à medida que atributos e julgadores são eliminados

3.4 ESTUDO DE CASO

O método sugerido é aplicado para selecionar atributos discriminantes e julgadores consistentes em um conjunto de dados sensoriais. Vinte e seis atributos sensoriais relacionados na Tabela 3.1 são avaliados por nove indivíduos treinados em um painel sensorial, o conjunto de dados está disponível mediante solicitação. Os atributos são relativos à aparência, sabor e textura. Os produtos analisados são diferentes formulações de cubos de carne ao molho, acondicionados em *pouches* termoestáveis. As avaliações são realizadas segundo o Método Spectrum, uma técnica Análise Descritiva Quantitativa (ADQ). No Método do Spectrum, os produtos são avaliados individualmente por cada membro do painel a respeito de um conjunto de atributos sensoriais [ver Meilgaard, Civille e Carr (1999) para a descrição do método]. Os produtos foram elaborados com base em especificações militares. Os painéis sensoriais foram realizados em 1994, no Nabisco Food Center da Rutgers University (EUA), como parte de um projeto de pesquisa para o exército norte-americano. Oito formulações foram testadas.

Nas análises a seguir, os atributos marcados com (*) não foram considerados devido à falta de observações de alguns julgadores. As avaliações foram feitas em quadruplicata com os vinte e quatro atributos. Assim, um total de 768 observações foram obtidas. A dimensão da matriz **X** contendo as avaliações de todos os julgadores é (288 x 24). Na análise o julgador *p* é referido como P*p*.

Tabela 3.1: Atributos sensoriais avaliados no experimento

Atributos de Aparência	**Atributos de Sabor**	**Atributos de Textura**
(1) Proporção de molho na carne	(4) Aroma de carne cozida	(15) Viscosidade do molho
(2) Espessura visual do molho	(5) Aroma de caldo de carne	(16) Elasticidade da carne
(*) Cor do molho (dados não avaliados)	(6) Aroma de carne cru	(17) Coesividade inicial da carne
(*) Cor dos cubos da carne (dados não avaliados)	(7) Aroma de proteína vegetal hidrolisada	(18) Densidade da carne
(3) Uniformidade de tamanho e forma da carne	(8) Carne com sangue coagulado	(19) Firmeza da carne
	(9) Espessura	(20) Maciez da carne
	(10) Aroma de carne queimada	(21) Fibrosidade da carne
	(11) Aroma de gordura	(22) Estratificação da carne
	(12) Sal	(23) Umidade da carne
	(13) Sensação metálica	(24) Película oleosa
	(14) Sensação de calor	

A análise de consistência dos julgadores é executada utilizando o índice de desempenho α, Eq. (1), como índice de consistência, a Tabela 3.2 mostra o resultado das leituras do alfa. O julgador 9 (P9) é o menos consistente e o primeiro a ser removido para seleção de atributos, seguido por P8 e similarmente daí em diante. Em geral, os participantes apresentam valores elevados de alfa, indicando consenso do grupo nas avaliações realizadas.

Tabela 3.2: Índice α de desempenho dos julgadores

ID do Julgador	Alfa
P1	0,8386
P2	0,8205
P3	0,8157
P4	0,8127
P5	0,8104
P6	0,8077
P7	0,7676
P8	0,6959
P9	0,0250

Na sequência, aplica-se PCA no conjunto de dados. Determinamos R=2 como o número de componentes a serem retidos em cada PCA realizada após remoção de um julgador. Este número de componentes explicam 63% ou mais da variância sobre os atributos, componentes adicionais explicam porções residuais da variância, não justificando assim a sua inclusão.

Quanto ao KKN, o parâmetro k = 3 foi definido através da validação cruzada. Inicialmente, um intervalo apropriado de k valores ímpares [1;11] e utiliza-se aquele k para classificar 80% do conjunto de dados, o k resultando a máxima acurácia da classificação é então usado para classificar os 20% restantes das observações. Esse procedimento é repetido inúmeras vezes, e o k que leva à maior acurácia média na porção dos 20% é escolhido.

A seleção dos atributos é realizada após a eliminação de cada julgador e conduz ao conjunto de perfis de acurácia dado na Fig. 3.2. A elimição dos julgadores é concluída quando um limite inferior de 5 julgadores restantes é alcançado. Pontos de fronteira da análise de PO são identificados no gráfico de acurácia. Esses pontos também são detalhados na Tabela 3.3.

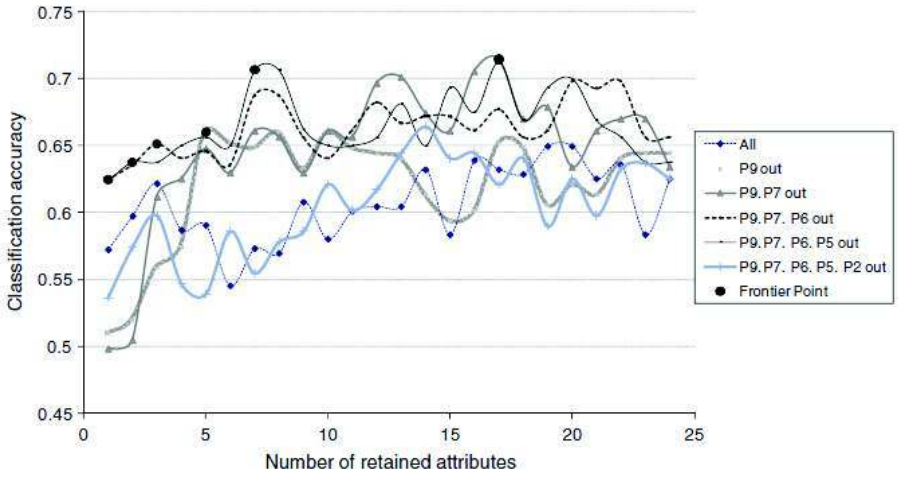


Figura 3.2: Perfil de acurácia à medida que atributos e julgadores são eliminados

O máximo de acurácia de 71,4% é obtida quando as avaliações dos julgadores P9, P7, P6 e P5 são removidas do conjunto de dados original, e 17 dos 24 atributos originais são mantidos. Isso leva a classificações 12% mais precisas do que usando KNN com todos os 24 atributos originais. Os atributos retidos são 20, 18, 16, 19, 22, 17, 4, 21, 3, 6, 23, 10, 12, 11, 1, 2 e 5, em ordem decrescente de relevância. Percebe-se que a acurácia da classificação utilizando 24 atributos e nove julgadores é menor do que a obtida quando 17 atributos e cinco julgadores são considerados. Isso é explicado pela remoção de atributos ruidosos e panelistas inconsistentes do conjunto de dados.

Soluções alternativas correspondem aos pontos de fronteira da Tabela 3.3, uma solução alternativa aparece claramente como promissora. FP5 apresenta uma situação em que 5 julgadores avaliam apenas 7 atributos e atingem uma acurácia de classificação de 70,6%. Reduzindo 70% do grupo original de atributos no experimento sensorial sem comprometer gravemente a acurácia da classificação é altamente desejável pois (*i*) o estresse dos julgadores é minimizado, e (*ii*) diminui o custo de execução da análise sensorial pelo painel.

Tabela 3.3: Informações sobre os pontos de fronteira do Pareto Ótimo

Pontos de Fronteira (FP)	Julgadores retidos	ID dos atributos retidos	Acurácia de classificação
1	P1, P2, P3, P4, P8	16	0,6245
2	P1, P2, P3, P4, P5, P8	16, 18	0,6375
3	P1, P2, P3, P4, P5, P6, P8	16, 19, 18	0,6510
4	P1, P2, P3, P4, P5, P6, P7, P8	22, 11, 21, 16, 18	0,7063
5	P1, P2, P3, P4, P8	20, 18, 16, 19, 22, 17, 4	0,7063
6	P1, P2, P3, P4, P8	20, 18, 16, 19, 22, 17, 4, 21, 3, 6, 23, 10, 12, 11, 1, 2, 8	0,7143

3.5 CONCLUSÃO

A redução do número de atributos a serem analisados em experimentos de perfil sensorial tem sido objeto de investigação nos últimos anos. O objetivo é identificar um subconjunto de atributos relevantes e não redundantes que permitam a discriminação das amostras analisadas no painel. A otimização é desejada para economizar tempo e fadiga aos julgadores, levando a uma coleta de dados menos onerosa. Muitas abordagens sobre seleção de atributos na literatura estão focadas na seleção dos atributos que responsáveis por grande quantidade de variação dos dados. O propósito deste artigo é duplo: (*i*) a identificar os atributos que conduzam a uma classificação dos produtos mais precisa e, (*ii*) que sejam informativos quanto a proporção da variância explicada no conjunto de dados multivariados.

Dessa forma, o método para selecionar os atributos importantes a serem utilizados na classificação de produtos avaliados por um painel sensorial consiste em: (1) Medir a consistência dos julgadores usando um índice apropriado, (2) Aplicar a PCA no conjunto de

dados consistindo de atributos sensoriais; (3) Computar um vetor de índices de importância de atributos baseada nos pesos da PCA, (4) Classificar o conjunto de dados sensoriais através da técnica dos *k*-vizinhos mais próximos (KNN) e calcular a acurácia da classificação. Iterar, eliminando o atributo com o menor índice de importância, classificando o conjunto de dados novamente, e re-computando a acurácia; (5) Construir um gráfico de acurácia, e (6) Remover o julgador menos consistente e executar uma nova seleção de atributos, repetindo os passos 2-6.

O método proposto foi aplicado em um conjunto de dados de análise descritiva composto de avaliações realizadas por 9 julgadores em 24 atributos, de 8 diferentes formulações de produtos. A máxima acurácia de 71,4% é obtida quando as avaliações de quatro membros do painel são removidas do conjunto de dados original, e 17 dos 24 atributos originais são retidos. Como alternativa, uma solução mais parcimoniosa é identificada através da análise de Pareto Ótimo onde cinco julgadores avaliam apenas 7 atributos e atingem uma acurácia de classificação de 70,6%. Futuras pesquisas incluem o desenvolvimento de alternativas para selecionar o melhor conjunto de atributos para classificação.

Apêndice

CÓDIGO DO MATLAB PARA SELEÇÃO DE ATRIBUTOS

```
function clas_PCA_KNN(ts,tr,ta,correct,Knn)
% tr=ts: evaluation data
% ta=correct: formulation classes
% Knn: number of nearest neighbors
[mtr,ntr]=size(tr);
tr1=zscore(tr);
ts1=zscore(ts);
 [coefs,scores,variances,t2] = princomp(tr1);
coefs;
WWW1=[sum(abs(coefs(:,1:2))')',(1:ntr)'];
WWW=flipud(sortrows(WWW1,1))
 tr1=tr1(:,WWW(:,2));
```

```
ts1=ts1(:,WWW(:,2));

WWW2=[WWW(:,2)];

for mm=0:ntr-1

    tr2=tr1(:,1:ntr-mm);

    ts2=ts1(:,1:ntr-mm);

    WWW22=WWW2(1:ntr-mm);

 train_patterns1=tr2';

test_patterns1=ts2';

train_targets=ta;

L= length(train_targets);

Uc= unique(train_targets);

 if (L < Knn),

   error('More neighbors than there are points.')

end

N = size(test_patterns1, 2);

test_targets = zeros(1,N);

for i = 1:N,

    dist=sum(((train_patterns1 - test_patterns1(:,i)*ones(1,L))).^2);

      [m, indices] = sort(dist);

        n = hist(train_targets(indices(1:Knn)), Uc);

        [m, best] = max(n);

        test_targets(i) = Uc(best);

end

G=test_targets;

H=G'==correct;

ACC_mm=sum(H)/mtr;

ACC(mm+1,1)=ACC_mm;

end

ACC
```

MATLAB CODES FOR PANELISTS' RANKING

```
function panelist(x, pain)

      %npain = number of painelists

      %X = evaluation data

[mx,nx]=size(x);

intpain=mx/npain

for i=1:npain;

    if i==1;

        x1=x(1:intpain:);

    elseif i==npain;

        x1=x((npain-1)*intpain)+1:npain*intpain,:);

    else i~=1 & i~=npain;

        x1=x(((1-i-1)*intpain)+1:i*intpain,:);

    end

SC=mean(x1);

[ysize,xsize]=size(SC);

  times=intpain;

for y=1:ysize,

for rep = 1:times,

out ((y-1)) * times + rep, :) = SC(y,:);

    end

    end

out1=out;

Xc=x1-out1;

t=trace(Xc'*Xc);

theta=1/sqrt(t);

Yc_i=theta*Xc;

YY(intpain*i+1:intpain*i+intepain,)=Yc_i;

end

      YY=YY(intpain+1:mx+intpain,:)
```

```
      S = xlswrite('tempdata.xls',YY)
function tpanelist(YY, tpain)
[mYY,nYY]=size(YY);
intpain=mYY/npain;
for k=1:npain
    if k==1;
        YY1=YY(1:intpain,:);
    elseif l==npain;
        YY2=YY((npain-1)*intpain)+1:npain*intpain,:);
    else l~=1 & l~=npain;
        YY2=YY(((l-1)*intpain)+1:i*intpain,:);
    end
s_l=(1+trace(YY1'*YY2))/2
ss(:,l+1)=s_l
end
sss(k+1,:)=ss
end
S=sss(2:k+1,2:l+1)
[V,D] = eig(S);
FEIG=V(:,npain);
FEIGnorm=FEIG/sum(FEIG)
for k=1:npain
    if k==1;
        YY1=YY(1:intpain,:);
    elseif k==npain;
        YY1=YY((npain-1)*intpain)+1:npain*intpain,:);
    else k~=1 & k~=npain;
        YY1=YY(((k-1)*intpain)+1:k*intpain,:);
    end
   C_k=FEIGnorm(k)*YY1;
```

```
    if k==1;
    CC=C_k;
     End
end
CCfor k=1:npain
    if k==1;
        YY1=YY(1:intpain,:);
    elseif k==npain;
        YY1=YY((npain-1)*intpain)+1:npain*intpain,:);
    else k~=1 & k~=npain;
        YY1=YY(((k-1)*intpain)+1:k*intpain,:);
    end
    alpha_k=trace(YY1'*CC)/srqt(trace(CC'*CC))
```

3.6 REFÊRENCIAS

ANZANELLO, M. J., ALBIN, S. L., & CHAOVALITWONGSE, W., Selecting the best variables for classifying production batches into two quality classes. **Chemometrics and Intelligent Laboratory Systems**, v.97(2), p.111–117, 2009.

AZAPAGIC, A., Life cycle assessment and its application to process selection, design and optimization. **Chemical Engineering Journal**, v.73(1), p.1–21, 1999.

CAMARA, J., ALVES, M., & MARQUES, J., Multivariate analysis for the classification and differentiation of Madeira wines according to the main grape varieties. **Talanta**, v.68, p.1512–1521, 2006.

CAPRON, X., SMEYERS-VERBEKE, J., & MASSART, D., Multivariate determination of the geographical origin of wines from four different countries. **Food Chemistry**, v.101, p.1585–1597, 2007.

CARBONELL, L., IZQUIERDO, L., & CARBONELL, I., Sensory analysis of Spanish mandarin juices: Selection of attributes and panel performance**. Food Quality and Preference**, v.18, p.329–34, 2007.

CHABANET, C., Statistical analysis of sensory profiling data. Graphs for presenting results (PCA and ANOVA). **Food Quality and Preference**, v.11(1–2), p.159–162, 2000.

CHAOVALITWONGSE, W., FAN, Y., & SACHDEO, C., On the time series k-nearest neighbor classification of abnormal brain activity. **IEEE Transactions on System and Man Cybernetics A**, v.37(6), p.1005–1016, 2007.

DEB, K., PRATAP, A., AGARWAL, S., & MEYARIVAN, T., A fast and elitist multi objective genetic algorithm: NSGA-II. **IEEE Transactions on Evolutionary Computation**, v.6(2), p.182–197, 2002.

DEB, K., THIELE, L., LAUMANNS, M., & ZITZLER, E., Scalable multi-objective optimization test problems. **Proceedings of the 2002 Congress on Evolutionary Computation**, v.1, p.825–830, 2002.

DERDE, M., & MASSART, D., Supervised pattern recognition: The ideal method? **Analytica Chimica Acta**, v.184, p.33–51, 1986.

DIJKSTERHUIS, G., FROST, M. B., & BYRNE, D. V., Selection of a subset of variables: Minimization of Procrustes loss between a subset and the full set. **Food Quality and Preference**, v.13, p.89–97, 2002.

DUDA, R., HART, P., & STORK, D., **Pattern Classification** (second ed.). New York: Wiley-Interscience, 2001.

FOGLIATTO, F. S., ALBIN, S. L., & TEPPER, B. J., A hierarchical approach to optimizing descriptive analysis multiresponse experiments. **Journal of Sensory Studies**, v.14(4), p.443–465, 1999.

GAUCHI, J., & CHAGNON, P., Comparison of selection methods of exploratory variables in PLS regression with application to manufacturing process data. **Chemometrics and Intelligent Laboratory Systems**, v.58, p.171–193, 2001.

GRANITTO, P., BIASIOLI, F., ENDRIZZI, I., & GASPERI, F., Discriminant models based on sensory evaluations: Single assessors versus panel average. **Food Quality and Preference,** v.19(6), p.589–595, 2008.

GRANITTO, P. M., GASPERI, F., BIASIOLI, F., TRAINOTTI, E., & FURLANELLO, C., Modern data mining tools in descriptive sensory analysis: A case study with a Random forest approach. **Food Quality and Preference**, v.18, p.681–689, 2007.

GUO, Q., WU, W., MASSART, D., BOUCON, C., & JONG, S., Feature selection in principal component analysis of analytical data. **Chemometrics and Intelligent Laboratory Systems**, v.61, p.123–132, 2002.

GUYON, I., & ELISSEEFF, A., An introduction to variable and feature selection. **Journal of Machine Learning Research**, v.3, p.1157–1182, 2003.

HORN, J., NAFPLIOTIS, N., & GOLDBERG, D., A niched pareto genetic algorithm for multiobjective optimization. In: Proceedings of the First IEEE Conference on Evolutionary Computation, **IEEE World Congress on Computational Intelligence**, v.1, p.82–87, 1994.

JANSEN, J., HOEFSLOOT, H., GREEF, J., TIMMERMAN, M., WESTERHUIS, J., & SMILDE, J., SCA: Analysis of multivariate data obtained from an experimental design. **Journal of Chemometrics**, v.19(9), p.469–481, 2006.

KUDO, M., & SKLANSKY, J., Comparison of algorithms that select features for pattern classifiers. **Pattern Recognition**, v.33, p.25–41, 2000.

LATREILLE, J., MAUGER, E., AMBROISINE, L., TENENHAUS, M., VINCENT, M., NAVARROC, S., Measurement of the reliability of sensory panel performances. **Food Quality and Preference**, v.17(5), p.369–375, 2006.

LAVIT, C., ESCOUFIER, Y., SABATIER, R., & TRAISSAC, P., The ACT (STATIS method). **Computational Statistics & Data Analysis**, v.18, p. 97–119, 1994.

LEDAUPHIN, S., HANAFI, M., & QANNARI, E. M., Assessment of the agreement among the subjects in fixed vocabulary profiling. **Food Quality and Preference**, v.17(3–4), p.277–280, 2006.

LIU, H., & YU, L., Toward integrating feature selection algorithms for classification and clustering. **IEEE Transactions on Knowledge and Data Engineering**, v.17(4), p.491–502, 2005.

LUCIANO, G., & NAES, T., Interpreting sensory data by combining principal component analysis and analysis of variance. **Food Quality and Preference**, v.20(3), p.167–175, 2009.

MALLET, Y., DE VEL, O., & COOMANS, D., Integrated feature extraction using adaptive wavelets. In: H. Liu & H. Motoda, F**eature extraction, construction and selection: A Data mining perspective**, p.175–189, 1998.

MARINI, F., BUCCI, R., MAGRI, A., & MAGRI, A., Authentication of Italian CDO wines by class-modeling techniques. **Chemometrics and Intelligent Laboratory Systems**, v.84, p.164–171, 2006.

MEILGAARD, M., CIVILLE, G. V., & CARR, B. T., **Sensory Evaluation Techniques** (3rd ed.). Boca Raton: CRC Press, 1999.

MONTGOMERY, D., PECK, E., & VINING, G., **Introduction to Linear Regression Analysis**. New York: John Wiley, 2001.

MURRAY, J. M., DELAHUNTY, C. M., & BAXTER, I. A., Descriptive sensory analysis: Past, present and future. **Food Research International**, v.34(6), p.461–471, 2001.

RASON, J., MARIN, J., DUFOUR, E., & LEBECQUE, A., Diversity of the sensory characteristics of traditional dry sausages from the centre of France. Relation with regional manufacturing practice. **Food Quality and Preference**, v.18(3), p.517–530, 2007.

REBOLO, S., PENA, R., LATORRE, M., BOTANA, A., & HERRERO, C., Characterisation of Galician (NW Spain) Ribeira Sacra wines using pattern recognition analysis. **Analytica Chimica Acta**, v.417, p.211–220, 2000.

RIPLEY, B., **Pattern Recognition and Neural Networks**. Cambridge: Cambridge University Press, 1996.

SAHMER, K., & QANNARI, E. M., Procedures for the selection of a subset of attributes in sensory profiling. **Food Quality and Preference**, v.19, p.141–145, 2008.

TABACHNICK, B. G., & FIDELL, L. S., **Using Multivariate Statistics**. New York: Harper Collins College Publishers, 1996.

TABOADA, H., & COIT, D., Data clustering of solutions for multiple objective system reliability optimization problems. **Quality Technology & Quantitative Management Journal**, v.4, p.35–54, 2007.

TABOADA, H., & COIT, D., Multi-objective scheduling problems: Determination of pruned Pareto sets. **IIE Transactions**, v.40, p.552–564, 2008.

URTUBIA, A., PERREZ-CORREA, J., SOTO, A., & PSZCZOLKOWSKI, P., Using data mining techniques to predict industrial wine problem fermentation. **Food Control,** v.18, p.1512–1517, 2007.

WEISS, S., APTE, C., DAMERAY, D., JOHNSON, D., PLES, F., GOETZ, T., *et al.,* Maximizing text-mining performance. **IEEE Intelligent Systems**, v.14(4), p.63–69, 1999.

WESTAD, F., HERSLETH, M., LEA, P., & MARTENS, H., Variable selection in PCA in sensory descriptive and consumer data. **Food Quality and Preference**, v.14, p.463–472, 2003.

WOLD, S., & SJOSTROM, M. A., Method for Analyzing Chemical Data in Terms of Similarity and Analogy (pp. 243–282). In B. R. Kowalski (Ed.), Chemometrics, Theory and Application, **ACS Symposium Series**. 52 (pp. 243–282). Washington, DC: American Chemical Society, 1977.

WU, X., KUMAR, V., QUINLAN, J., GHOSH, J., YANG, Q., MOTODA, H., *et al.* Top 10 algorithms in data mining. **Knowledge and Information Systems**, v.14(1), p.1–37, 2008.

ZITZLER, E., & THIELE, L., **Multi objective evolutionary algorithms: A comparative case study and the strength Pareto**, 1999.

4 ARTIGO 3 – PLS discriminant analysis applied to conventional sensory profiling data

Karina Rossini

Stephane Verdun

Veronique Cariou

El Mostafa Qannari

Flávio Sanson Fogliatto

Artigo publicado no Vol. 23 da revista Food Quality and Preference (ISSN 0950-3293)

Abstract

Several methods have been proposed in the literature to analyse conventional sensory profiling data. We focus on factor analytical methods which have been extensively used due to their ability to produce graphical displays which are both useful and easy to interpret. Available factor analytical methods include principal components analysis on averaged assessors data or on the data matrix obtained by stacking the assessors' datasets one on top of the others, and canonical variates analysis; each method has advantages and drawbacks. As an alternative it is advocated the use of *PLS* discriminant analysis, which is at the intersection of the methods mentioned above. It provides statistical tools to assess on the one hand the agreement among assessors and the discrimination among products by means of the between to total variance ratio, and on the other hand the relative importance of variables by means of *VIP* (Variable Importance in the Projection) indices. The VIP indices may also be useful to guide the selection of a subset of relevant attributes from the complete set of attributes. In this paper, *PLS* discriminant analysis is compared with other methods, and results are illustrated through a case study. In particular, the stability of the various methods is investigated using assessors' re-sampling (bootstrap) and confidence ellipses.

Key words: PLS discriminant analysis, Principal Components Analysis, Canonical Variates Analysis, Sensory profiling data, Selection of variables, Bootstrap, Confidence ellipses.

4.1 INTRODUCTION

The main goal of the statistical treatment of conventional sensory profiling data is to exhibit inter-product differences while handling the variations among assessors. That is often accomplished by means of factor analytical methods which enable the investigation of similarities between products on the basis of graphical displays. Different approaches have been proposed to address this issue; among them we single out Principal Components Analysis (*PCA*), Generalized Procrustes Analysis (*GPA*), and Canonical Variates Analysis (*CVA*). Each approach has advantages and drawbacks. We focus on conventional sensory data, also referred to as sensory data with a fixed vocabulary. This type of data may be obtained by means of different sensory evaluation procedures (e.g. the Quantitative Descriptive Analysis protocol by Stone and Sidel, 1998), and may be organized as a three way data matrix (products × attributes × assessors). However, depending on the statistical method to be performed on the data, other data arrangements may be more convenient.

PLS Discriminant Analysis (*PLS-DA*) is yet another method which is suited for the analysis of conventional sensory profiling, and stands at the intersection of the methods mentioned above. It yields interesting indices, such as the between to total variance ratios, which reflect the agreement among assessors and the discrimination among products. The *VIP* indices (Variable Importance in the Projection) are also of paramount interest as they highlight the importance of the various attributes. They may be useful in guiding the selection of a subset of relevant attributes from the complete set of attributes. We also show how the graphical displays of the products may be enhanced by using confidence ellipses obtained by means of assessors' re-sampling (bootstrap). The outcomes of *PLS-DA* are compared to those of alternative methods through a case study pertaining to the sensory evaluation of varieties of cider.

The rest of the paper is organized as follows. We start by discussing the pre-treatment of the data in order to cope with some known sources of variation among assessors. Next we outline the most popular methods to analyse conventional sensory profiling data and focus on *PLS-DA*, presenting its advantages in comparison with other methods. In particular, we discuss the *VIP* indices which are popular within the framework of *PLS* regression and

PLS-DA, and present how they may be a useful guiding criterion to select a subset of relevant attributes from a complete set. Finally, comparison of *PLS* discriminant analysis with other methods is illustrated using a case study dataset. In particular, the various methods' stability is investigated using assessors' re-sampling (bootstrap) and confidence ellipses.

4.2 MATERIAL AND METHODS

4.2.1 Pre-treatment of the data

Assume that *m* assessors perform the sensory profiling of *n* products using a set of *p* attributes. Data obtained from assessor *k* (=1,…,*m*) is organized in a (*n* × *p*) matrix X_k^*, with rows referring to products and columns referring to attributes. To cope with some known sources of variation among assessors it is recommended to centre each matrix X_k^* by subtracting from the entries in each column its corresponding average. Centring removes the assessors' main effect (or shift effect) which will be present if assessors use different levels of the scoring scale. Another source of variation is associated with assessors using different ranges of the scoring scale. Isotropic scaling factors are usually introduced to address this problem, as follows. Multiply each dataset X_k^* by a scaling factor α_k to (*i*) shrink configurations of assessors with tendency to use large ranges of the scoring scale (for that, make $\alpha_k < 1$), or (*ii*) expand configurations of assessors with tendency to use relatively narrow ranges of the scoring scale (for that, make $\alpha_k > 1$). Appropriate scaling factors may be computed as follows (Kunert and Qannari, 1999):

- Determine t_k, the total variance of the dataset X_k^*, by adding the variances of each column of X_k^*;
- Compute t as the average of t_k (*k*=1,…,*m*);
- Set $\alpha_k = \sqrt{\frac{t}{t_k}}$.

Multiplying each dataset X_k^* by its associated scaling factor α_k yields new matrices with the same total variance, *t*. In what follows, we denote by X_k the (*n* × *p*) matrix obtained

by centring the columns of X_k^* and multiplying the resulting matrix by the isotropic scaling factor α_k.

4.2.2 Statistical treatments of sensory profiling data

There are different ways to organize the data obtained from a conventional sensory profiling procedure. Certain statistical analyses follow naturally from each choice of arrangement, as depicted in Figure 4.1.

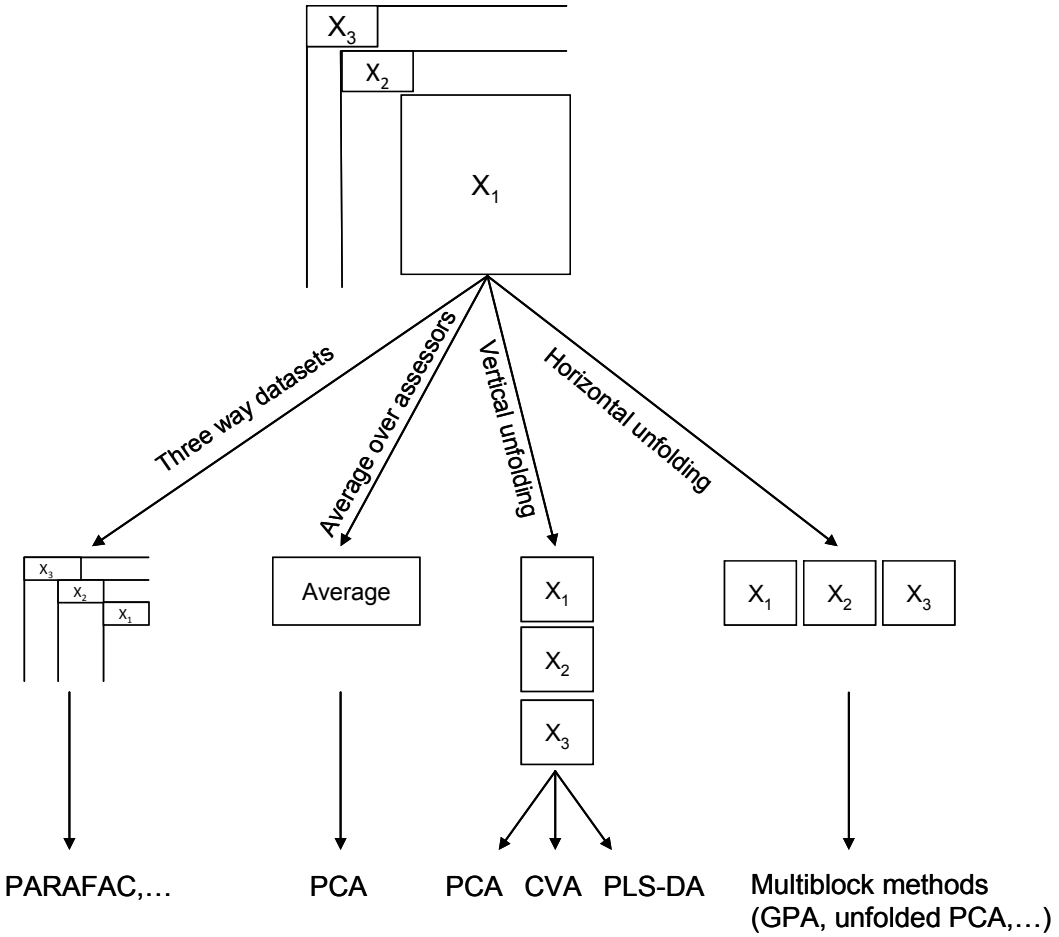


Figura 4.1: Conventional sensory data arrangements leading to different methods of statistical treatment

As stated previously, sensory profiling data may be presented as a three way array in which entries refer to products, attributes and assessors, and methods devoted to this type of arrangement (e.g. *PARAFAC*) may be performed on the data (Brockhoff, Hirst and Næs, 1996; Cocchi *et al.*, 2006; Bro *et al.*, 2008). Such methods are not very popular in sensory

evaluation probably because their rationale is not fully grasped by practitioners. Notwithstanding, we believe that they should be further investigated by sensory analysts.

Average dataset. The most popular practice in sensory analysis is the averaging of datasets over assessors, resulting in a two way dataset (products × attributes). Obviously, the rationale behind such practice is to consider evaluations given by assessors as replicates which differ from each other by random noise. *PCA* is usually performed on the average dataset in order to depict the relationships among products. However, this strategy of analysis does not explore the within-products variation structure, and unless further investigation is carried out it does not provide tools, such as indices and graphical displays, to assess differences among the assessors.

Horizontally unfolded data. In this case each dataset X_k is placed sideways horizontally, and the resulting supermatrix X^H is suitable to be analyzed using methods developed within the framework of multiblock datasets. The assumption is that although using the same attributes, assessors might interpret them differently, and analytical methods that adjust for this source of variation should be considered. Undoubtedly, the most popular among such methods is Generalized Procrustes Analysis (*GPA*) (Gower, 1975; Dijksterhuis and Gower, 1991). Alternatively, a family of methods based on performing *PCA* on X^H may also be used; Multiple Factor Analysis (*MFA*) (Husson, Le Dien and Pagès, 2005) and *STATIS* (*Structuration des Tableaux à Trois Indices de la Statistique*) (Schlich, 1996) are examples of such methods. However, in the sections to follow, these methods are not considered since we are concerned with genuine situations of conventional sensory profiling, where the assumption is that assessors interpret attributes similarly.

Vertically unfolded data. In this case datasets X_k are stacked up vertically; the resulting supermatrix X^V has ($n \times m$) rows and p columns. *PCA* may be performed on X^V and we refer to Luciano and Naes (2009) for an interesting discussion on how the *PCA* outcomes could be used in conjunction with *ANOVA* to investigate both the sensory data structure and the similarities among products. For the same analytical purpose, Monrozier and Danzart (2001) among others propose the use of Canonical Variate Analysis (*CVA*). In *CVA* the number of groups is equal to the number of products, and we seek components (also called canonical variates) that best discriminate the products. As a result, we want products as far removed from each other as possible (maximizing the variation between groups) and observations within groups (i.e. products) as much clustered around their centroids as possible

(minimizing the variation within groups). *CVA* is clearly more aligned with the objectives of sensory profiling. However, it is well known that, similarly to multiple linear regression, *CVA* may lead to unstable results in the presence of high collinearity among attributes; *PLS-DA* is recommended to overcome such drawback (Naes and Indahl, 1998; Barker and Rayens, 2003; Nocairi, Qannari, Vigneau and Bertrand, 2005). Martens and Martens (2001) were the first to propose the use of *PLS-DA* on sensory profiling data but, to the best of our knowledge, this paper is the first to elaborate on the topic.

4.2.3 PLS-DA applied to conventional sensory profiling

In short, *PLS-DA* seeks to determine components, also called latent variables, which maximize the variation between groups (products); see Barker and Rayens (2003), and Noicari *et al.* (2005). From this standpoint, *PLS-DA* appears to be at the intersection of the following three analytical approaches:

(*i*) *PCA* applied on the average dataset − the objective is to recover the total variance in the sensory data averaged over all assessors (i.e. the between-products variation), but assessors' individual data are not taken into account;

(*ii*) *PCA* applied on matrix X^V − the objective is to recover the total variance from all datasets, but the presence of groups (products) in the data is not explicitly taken into account in the computation of the principal components;

(*iii*) *CVA* − the objective is to recover the between-products variation while minimizing the within-products variation; however, as stated above, this method is sensitive to the presence of collinearity among attributes.

PLS-DA yields graphical displays and offers interpretation tools that enable investigating the structure of the sensory data. *PLS-DA* components may be used to represent the rows in matrix X^V. We may also represent on the basis of the same components any given product as the centroid (or average) point of the rows corresponding to this product; that is, the point which stands at the barycentre of the assessors' observations corresponding to this product. The number of significant components to be retained is usually determined by a cross-validation procedure (Martens and Naes, 1989). Moreover, *ANOVA*s may be performed on *PLS-DA* components rather than on principal components extracted from matrix X^V, as proposed by Luciano and Naes (2009). *PLS-DA* components are likely to offer a better

discrimination of products if compared to *PCA* components. We may also compute a discrimination index for each component, given by the ratio of the between-products variance (i.e. the variance of the averaged scores over all assessors) to the total variance. The same calculation procedure may be extended to components derived through *PCA* performed on X^V and through *CVA*. In the case of principal components derived from *PCA* on the average dataset, the discrimination indices correspond to the same ratio above, with variances obtained as follows. The between-products variance is equivalent to the variance of the principal component scores. To obtain the total variance, individual data from each assessor are superimposed on the principal component under consideration (i.e. the same vector of loadings associated with this principal component is applied to the individual datasets); the variance of the superimposed scores is used as an estimate of the total variance.

From a technical standpoint, PLS-DA seeks, step by step, latent variables (or components) which are linear combinations of the columns in X^V. Let us denote by $t = X^V a$ the first latent variable, where a is the vector of loadings constrained to unit lenght. As mentioned above, t is sought such that the between groups (or products) variance is as large as possible. The solution to this problem leads to set a as the eigenvector of the between groups variance-covariance matrix associated with the largest eigenvalue. Thereafter, the so-called deflation procedure is applied in order to determine a second latent variable. This consists in regressing all the variables in X^V upon t and considering the dataset formed by the residuals. Then, taking this latter dataset instead of X^V the same procedure aiming at maximizing the between groups variance is again performed, thus leading to a new latent variable which is by construction orthogonal to t. The same strategy can be reiterated in order to determine subsequent latent variables.

Within the context of *PLS* Regression and *PLS-DA*, the *VIP* (variable importance in the projection) indices are of paramount interest (Chong and Jun, 2005). *VIPs* are associated with attributes, and reflect their contribution in discriminating the products. They may be computed for each component separately or they may be computed for a *PLS-DA* model which includes several components. As a rule of thumb proposed by Wold (1994), we may discard from the model attributes with small *VIP* values (smaller than 0.8). Therefore, as a strategy to select a subset of attributes from a larger set one may set up a model with an appropriate number of components, and discard attributes with small *VIP* values (Chong and Jun, 2005).

4.2.4 Confidence ellipses

In sensory profiling analysis, Husson and Pagès (2005) have stressed the benefits of setting up confidence ellipses on the graphical displays that depict the similarity between products. That enables identifying products that are significantly different from the group in a multivariate setting. The authors propose a bootstrapping approach to set up the confidence ellipses. The approach relies on intensive computation since it generates a large number of (virtual) panels, which are in turn submitted to *PLS-DA*. As a result, it is possible to assess fluctuations in the position of the products on the graphical displays. Ideally, these fluctuations should be small, reflecting a good stability of the model under consideration.

The resampling (bootstrap) strategy may be implemented in four steps:

Step 1. Perform a *PLS-DA* on the sensory data obtained from *m* assessors. Assume that a *PLS-DA* model with *r* components is retained. As stated above, these components may be used to depict relationships between products.

Step 2. Perform a resampling from the *m* assessors. In other words, create a new panel comprised of *m* assessors by randomly selecting assessors from the initial panel, with replacement. Consequently, in a bootstrapped panel a given assessor may be selected several times whereas others may never be selected.

Step 3. Perform a *PLS-DA* on data from the bootstrapped panel and retain a model with *r* components. Next, the new positions of the products are adjusted to the original positions obtained in Step 1 by means of a procrustean rotation (see, for instance, Krzanowski, 2000).

Step 4. Reiterate steps 2 and 3 a large number of times (say, 5000 times). Eventually, for each product, a confidence ellipse containing a desired percentage (e.g. 95%) of the resampled points associated with this product may be drawn.

We use the bootstrap resampling technique described above to compare several methods for the analysis of conventional sensory profiling data; namely: *PCA* on the average dataset, *PCA* on the supermatrix X^V, *CVA*, and *PLS-DA*.

4.3 CASE STUDY

To illustrate the use of *PLS-DA* in conventional sensory profiling and compare its outcomes with those obtained using other methods of analysis, we consider a case study in which a quantitative descriptive analysis is performed on ten varieties of cider. The sensory panel is formed by seven trained assessors who were asked to score products using a list of ten sensory attributes: sweet, intensity of odour, acid, bitter, astringency, strength, pungent, alcohol, perfume and fruity.

4.4 RESULTS AND DISCUSSION

4.4.1 Pre-treatment of sensory data

Table 1 shows the isotropic scaling factors which were applied to the datasets associated with the seven assessors. Isotropic scaling factors smaller than 1 indicate that configurations of the associated assessors were shrunk to correct their tendency to use a relatively large range of the scale; that corresponds to assessors 4 and 5 in Table 4.1. In opposition, configurations of assessors 1, 2, 6, and to a lesser extent assessors 3 and 7, were expanded since their associated scaling factors are larger than 1.

Table 4.1: Isotropic scaling factors used in the pretreatment of the sensory data

Assessor	1	2	3	4	5	6	7
Isotropic scaling factor pretreatment	1.207	1.117	1.020	0.737	0.819	1.854	1.087

4.4.2 Discrimination indices

PLS-DA was performed on the cider data. In a cross-validation procedure only the first three components turned out significant. To assess the extent to which products are discriminated by the three retained *PLS-DA* components, we propose computing for each component a discrimination index given by the between-products variance to total variance ratio.

Table 4.2 gives the discrimination indices associated with the first three *PLS-DA* components. To allow comparisons, we also present the discrimination indices obtained through *CVA*, *PCA* performed on X^V and *PCA* performed on the average dataset. Not

surprisingly, *CVA* gives the largest discrimination indices. In fact, as stated above, *CVA* aims at maximizing the between-products variance and minimizing the within-products variance, and we can show that this is equivalent to maximizing the discrimination index (between-products variance to total variance ratio).

For the first component, *PLS-DA* and *PCA* on the average dataset have the same performance in terms of discrimination. In general, *PLS-DA* leads to better results than the other methods, except for *CVA*.

Tabela 4.2: Discriminant indices for the first three components derived from PLS-DA, CVA and PCA

Method	Axis 1	Axis 2	Axis 3
PLS-DA	0.887	0.701	0.417
CVA	0.927	0.781	0.430
PCA on the average data set	0.886	0.698	0.410
PCA on the concatenated datasets	0.868	0.600	0.274

4.4.3 Graphical displays

Figure 4.2 displays the biplots where both variables and products are plotted on the same plane. For simplicity, we have retained only configurations on the first factorial plan, comprised of the first two retained components. To allow comparisons, we also present the biplots obtained through *CVA*, and through *PCA* on the average dataset and on the dataset obtained by stacking the assessors' datasets one on top of the others. It is clear that the four configurations lead to similar results. The first *PLS-DA* component (axis 1) opposes ciders 4, 8 and 10 to ciders 7, 5, 2, and 6. The former group of ciders is sweeter, fruitier, more

perfumed, less bitter, less strong and less pungent than the second group. The second component (axis 2) mainly singles out cider 9 which has a more intense odour.

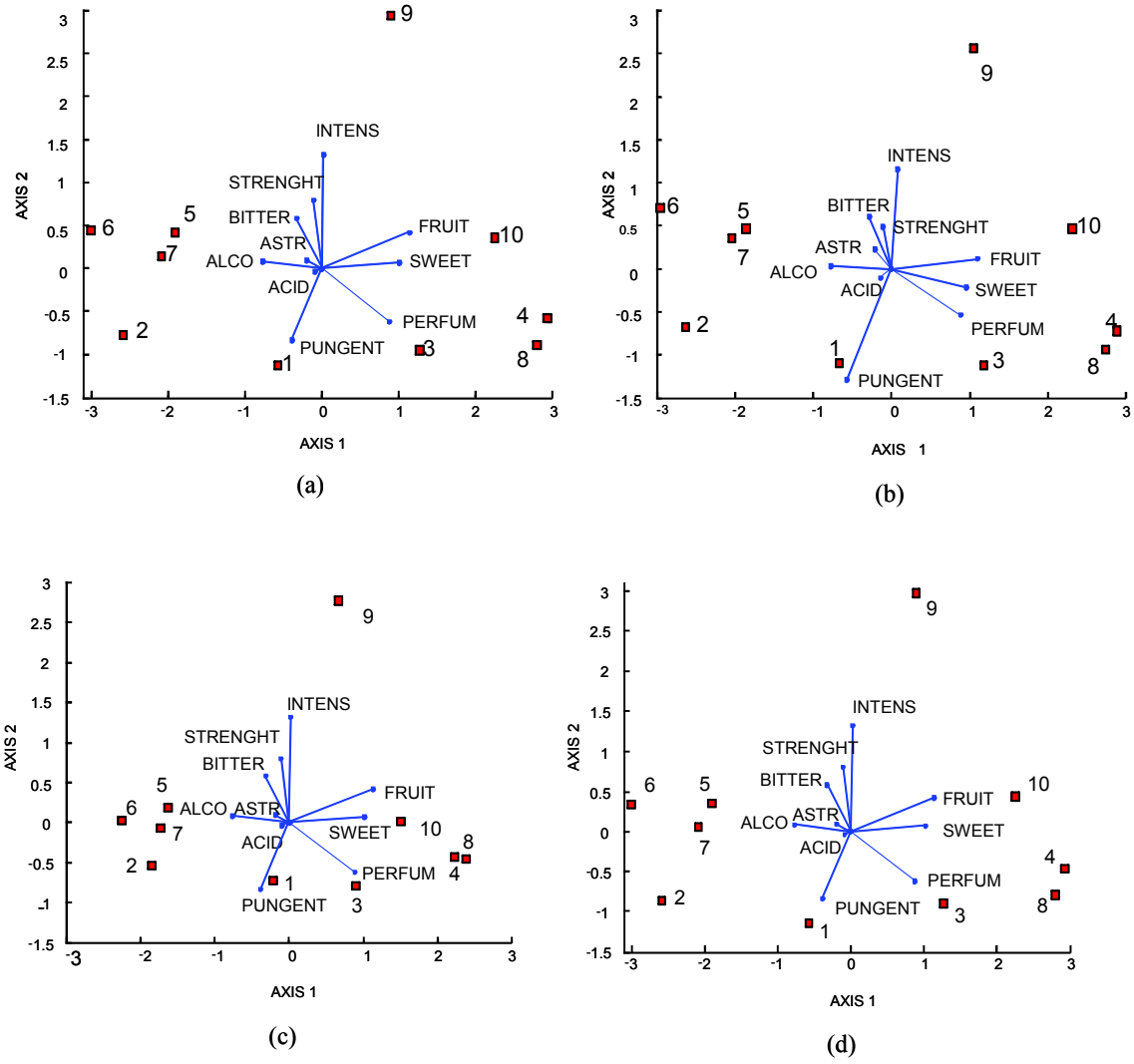


Figure 4.2: Configuration of products on the first factorial plan obtained using (a) PLS-DA, (b) CVA, (c) PCA on the average dataset and (d) PCA on the concatenated dataset

4.4.4 Confidence ellipses around the products

Figure 4.3 presents the 95% confidence ellipses around the products, which were obtained using the bootstrapping procedure in Section 2.4. The ellipses reflect the variability of the sensory evaluations and the discrimination of products in a multivariate setting. Four sets of ellipses are displayed, one for each analytical approach. Regardless of the method used to analyse the sensory data, products 1, 3 and 9 are clearly separated from the others. Product

10 is separated from the other products, except in the graphical display obtained through *CVA*. *PLS-DA* provides a good separation of products 2 and 6, while for the other methods of analysis these products' confidence ellipses overlap with those of other products. In general, it is clear that *PLS-DA* best discriminates products since their corresponding confidence ellipses are smaller than those obtained by alternative methods. In addition, it is noteworthy that *CVA* which is by principle focused on the discrimination of products, generates confidence ellipses that overlap the most. That is probably due to the method's instability when collinearity among attributes is present in the dataset.

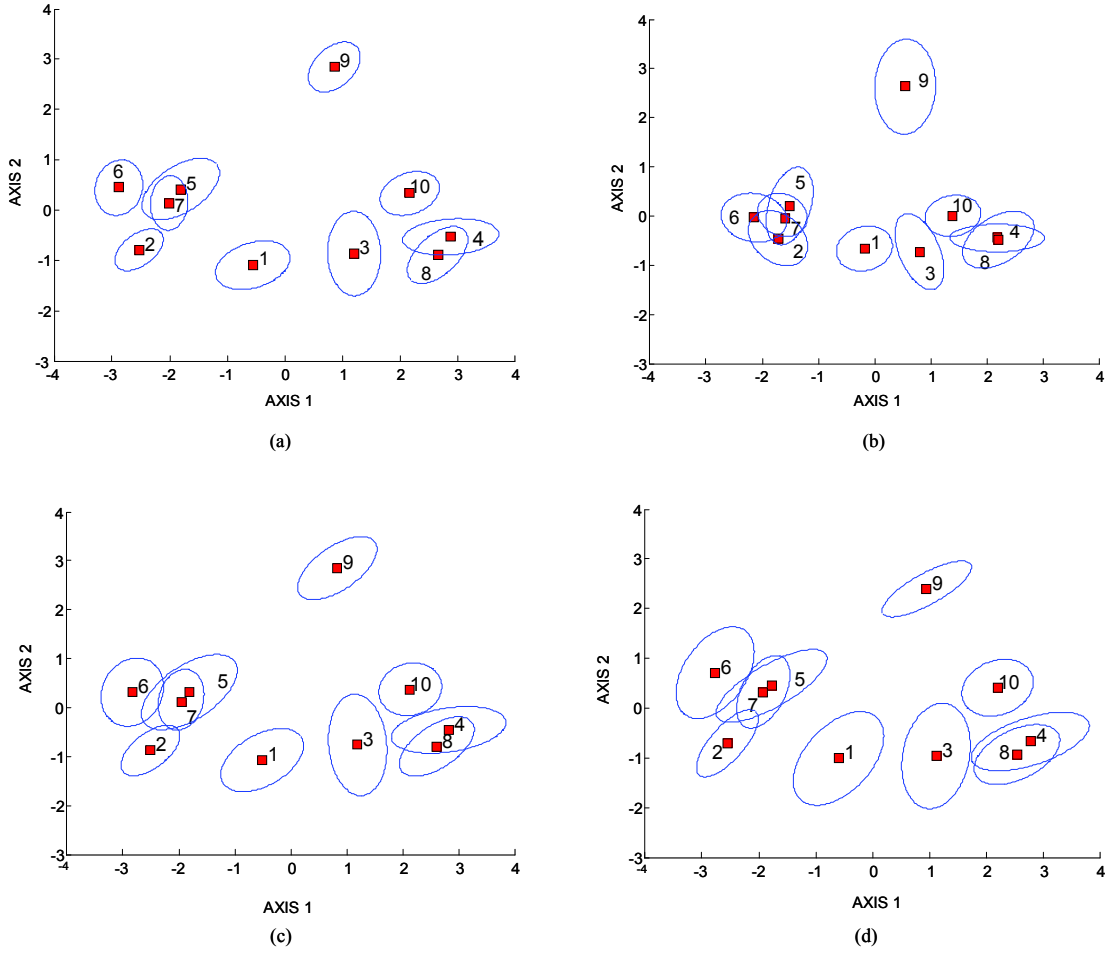


Figure 4.3: Configuration of products on the first factorial plan obtained using (a) PLS-DA, (b) CVA, (c) PCA on the average dataset, and (d) PCA on the concatenated dataset and their corresponding 95% confidence ellipses

4.4.5 VIP indices and selection of a subset of attributes

Figure 4.4 presents the *VIP* indices associated with a *PLS-DA* model comprised of three components, and their 95% confidence intervals obtained by means of the bootstrap procedure in Section 2.4. It highlights the importance of the attributes fruity, perfume, intensity of odour and sweet. The least important attributes are acid, bitter astringent and strength of odour which have *VIP* indices smaller than 0.8 or, considering the confidence intervals, not significantly larger than 0.8; in subsequent studies these attributes could be discarded thus saving time and fatigue to assessors. For completeness we have discarded these attributes and performed a *PLS-DA* on the remaining variables. Figure 4.5 presents the configuration of products on the first factorial plane corresponding to the first two *PLS-DA* components, as well as the 95% confidence ellipses obtained by means of bootstrap resampling. As expected, the overall structure of product configuration is preserved; i.e. disregarding attributes with low *VIP* values does not compromise the characterization of products.

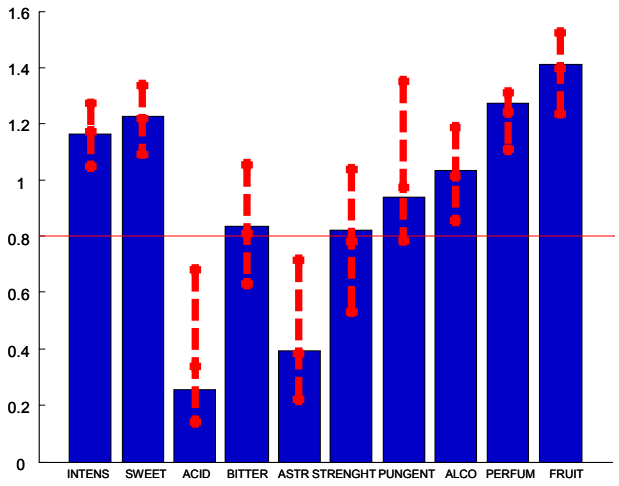


Figure 4.4: VIP indices with 95% bootstrap confidence intervals for a PLS-DA model comprised of three components

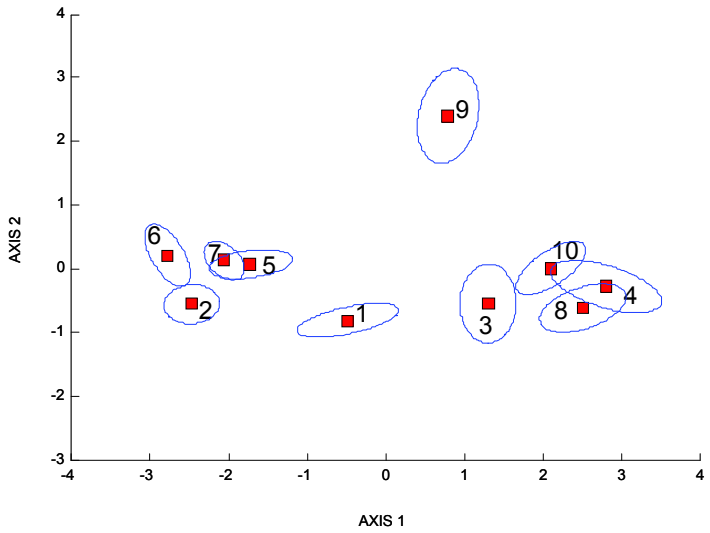


Figure 4.5: Configuration of products and 95% confidence ellipses obtained by means of PLS-DA and bootstrap resampling performed on the subset of attributes with VIP indices higher that 0.8

4.5 CONCLUSION

In this paper we propose the use of *PLS-DA* for the analysis of conventional sensory profiling. *PLS-DA* stands at the crossroads of popular methods of statistical treatment of conventional sensory profiling data, namely *PCA* on the concatenated dataset, *PCA* on average dataset and *CVA*. It provides statistical tools which allow a better interpretation of the analytical outcomes; among them we single out graphical displays, discrimination indices and *VIP* indices. The last two indices reflect the importance of variables in the discrimination and may be used within a strategy to select a subset of attributes from a larger set. In addition, assessors re-sampling (bootstrap) techniques provide tools which enable checking the stability of outcomes and the discrimination of products in a multivariate setting. In the light of a case study presented in the paper it turned out that *PLS-DA* provides better stability and achieves better discrimination of products than competing methods.

Future research will be deployed in the following directions: (*i*) deeper analysis on the merits of *PLS-DA* over alternative methods; (*ii*) broader investigation on the proposed strategy to select a subset of attributes, and comparison with other variable selection methods (as reviewed in Sahmer and Qannari, 2008); and (*iii*) proposition of a method to directly relate *PLS-DA* outcomes to preference data, when such data are available. Research deployment (*iii*) is particularly challenging since it consists of relating preference data to sensory data, while taking account of the individual assessors' evaluations instead of merely considering averaged scores, as it is usually done.

4.6 REFERENCES

BARKER, M. AND RAYENS, W., Partial Least Squares for discrimination. **Journal of Chemometrics**, v.17(3), p. 166-173, 2003.

BRO, R., QANNARI, E. M., KIERS, H.A.L., NÆS, T. AND FRØST M. B., Multiway models for sensory profiling data. **Journal of Chemometrics**, v.22(1), p. 36-45, 2008.

BROCKHOFF, P.M., HIRST, D. AND NÆS, T., **Analyzing individual profiles by three-way**, 1996.

CHONG, IL-GYO AND JUN, CHI-HYUCK, Performance of some variable selection methods when multicollinearity is present. **Chemometrics and Intelligent Laboratory Systems,** v.78(1-2), p. 103–112, 2005.

COCCHI, M., BRO, R., DURANTE, C., MANZINI, D., MARCHETTI, A., SACCANI, F., SIGHINOLFI, S. AND ULRICI, A., Analysis of sensory data of Aceto Balsamico Tradizionale di Modena (ABTM) of different ageing by application of PARAFAC models. **Food Quality and Preference**, v.17(6), p. 419-428, 2006.

DIJKSTERHUIS, G. B. AND GOWER, J. C., The interpretation of Generalized Procrustes analysis and allied methods. **Food Quality and Preference**, v.3(2), p. 67-87, 1991.

T. Naes and E. Risvik (Eds), Multivariate Analysis of factor analysis. In: **Data in Sensory Science.** Elsevier Science Publishers.

GOWER J. C., Generalized Procrustes analysis. **Psychometrika**, v.40(1), p.33-51, 1975.

HUSSON F., LÊ, D. S. AND PAGÈS, J., Confidence ellipse for the sensory profiles obtained by Principal Component Analysis. **Food Quality and Preference**, v.16(3), p.245-250, 2005.

KRZANOWSKI, W. J., **Principles of Multivariate Analysis**; a User's Perspective, Clarendon Press, Oxford, 2000.

KUNERT, J. AND QANNARI, E. M., A Simple alternative to generalized Procrustes Analysis: Application to Sensory Profiling Data. **Journal of Sensory Studies**, v.14(2), p.197-208, 1999.

LUCIANO, G. AND NAES, T., Interpreting sensory data by combining principal component analysis and analysis of variance. **Food Quality and Preference**, v.20(3), p.167-175, 2009.

MARTENS H. AND NAES T. **Multivariate Calibration**. Chichester etc.: Wiley, 1989.

MARTENS, H., AND MARTENS, M. **Multivariate analysis of quality: An introduction**. UK: Wiley Chichester, 2001.

MONROZIER R. AND DANZART M., A quality measurement for sensory profile analysis: the contribution of extended cross-validation and resampling techniques. **Food Quality and Preference**, v.12(5-7), p.393-406, 2001.

NAES, T. AND INDAHL, U., A unified description of classical classification methods for multicollinear data. **Journal of Chemometrics**, v.12(3), p.205-220, 1998.

NOCAIRI, H., QANNARI, E. M., VIGNEAU, E. AND BERTRAND D., Discrimination on latent variables with respect to patterns with application to multi-collinear data. **Computational Statistics and Data Analysis**, v.48(1), p.139-147, 2005.

SAHMER K.AND. QANNARI E.M., Procedures for the selection of a subset of attributes in sensory profiling. **Food Quality and Preference**, v.19(2), p.141-145, 2008.

SCHLICH P., Defining and validating assessor compromises about product distances and attribute correlations. In T. Naes and E. Risvik (Eds.) **Multivariate analysis of data in sensory science**. Elsevier Science Publishers. 1996.

STONE, H. AND SIDEL, J. L., Quantitative descriptive analysis: developments, applications, and the future. **Food Technology Journal**, v.52, p.48–52, 1998.

WOLD, S., JOHANSSON, E. AND COCCHI, M., 3D **QSAR in Drug Design; Theory, Methods, and Applications**, ESCOM, Leiden, Holland, p.523– 550, 1994.

5 ARTIGO 4 – Método de seleção de variáveis para minimização da variância de predição

Karina Rossini

Veronique Cariou

El Mostafa Qannari

Flávio Sanson Fogliatto

Resumo

A seleção de variáveis é um tema de bastante interesse em bancos de dados de espectros de infravermelho (NIR), que podem apresentar milhares de variáveis. O processo de seleção visa elimina as variáveis irrelevantes e de ruído, gerando um modelo de regressão constituído do subconjunto relevante de variáveis. Este artigo propõe um método para selecionar variáveis maximizando a correlação entre as variáveis selecionadas e a variável de resposta, selecionando variáveis independentes e não colineares. Além disso, utiliza o Critério de Informação de Akaike como critério de parada. O método é aplicado a dois bancos de dados experimentais, com medições de amostras de damasco e milho em espectro infravermelho, e os resultados são comparados com outro método disponível na literatura. Os resultados são promissores: o método proposto retém um menor número de variáveis, selecionando às que fornecem modelos de regressão com melhores adequações de ajuste para fins de predição.

Palavras-chave: seleção de variáveis, espectro infravermelho, CovSel

Abstract

Variable selection is a relevant topic among researchers dealing with near-infrared (NIR) spectrum datasets, which may present thousands of variables. The selection process is aimed at discarding irrelevant and noisy variables, leading to regression models comprised of a subset of relevant variables. In this article we propose a variable selection method in which the objective is to maximize the correlation between independent and response variables, while selecting independent variables that are not collinear. In addition, we propose using the Akaike Information Criterion as stopping criterion in the selection process. The method is applied to two experimental datasets, comprised of NIR spectra measured in apricot and corn samples, and the results are compared with another method available in the literature. Results

were promising: the proposed method led to a smaller number of variables selected, choosing those that produced regression models with better adequacy of fit.

Key words: Variable selection, near infrared spectrometry, CovSel

5.1 INTRODUÇÃO

Em estudos de quimiometria e química, observações acerca de um grande número de variáveis são coletadas em relação ao número de amostras. Nestes estudos, dados mal condicionados podem apresentar problemas quando métodos como a regressão linear múltipla (MLR) são utilizados, devido ao que se denomina colinearidade dos dados. Métodos alternativos, tais como regressão dos mínimos quadrados parciais (PLS), podem lidar com este problema devido à sua robustez no tratamento dos dados.

Através de um procedimento iterativo o qual gera combinações lineares entre as variáveis de processo e as variáveis de resposta, a regressão PLS visa maximizar a covariância entre dois conjuntos de variáveis, fornecendo parâmetros com informações úteis para identificar variáveis de processo que explicam a maior parte da variabilidade de resposta. Wu e Manne (2000), Wold *et al.* (2001a e 2001b), Muteki e MacGregor (2007), Lee *et al.* (2009) e Kohonen *et al.* (2009) fornecem detalhes sobre regressão PLS; uma discussão comparativa sobre as propriedades do PLS e MLR é relatado por Almoy (1996) e Dingstad *et al.* (2004).

A seleção de um subconjunto de variáveis relevantes da totalidade das variáveis candidatas tem sido uma preocupação constante na literatura. Uma variedade de métodos tem sido proposta com base em paradigmas diversos, tais como procedimentos sequenciais (Anzanello *et al.*, 2011), Baumann *et al.* (2002), Roy e Roy (2008), análise de agrupamentos (Sahmer e Qannari, 2008) e algoritmo genético (Ferrand *et al.*, 2010, e Leardi *et al.*, 2002).

A regressão PLS identifica a relação entre variáveis irrelevantes e de ruído, gerando um modelo de regressão constituído do subconjunto relevante de variáveis (Gauchi e Chagnon, 2001). Neste contexto, diversas aplicações estão descritas na literatura: Wang *et al.* (2006) em processos de refinação, Kohonen *et al.* (2009) na indústria de fertilizantes, Wold *et al.* (2001a) em reciclagem de papel, Hoskuldsson (2001), Xu *et al.* (2007) e Xiaobo *et al.* (2010) em espectro de infravermelho (NIR) de dados, e Baumann *et al.* (2002), Zhai *et al.* (2006), Xu *et al.* (2008) e Deeb (2010) em relação estrutura-atividade quantitativa / relação estrutura-propriedade quantitativa (QSAR / QSPR). A regressão PLS também foi combinada

com uma grande variedade de técnicas de análise multivariada e de otimização para seleção de variáveis, como relatado por Gauchi e Chagnon (2001), Xu e Zhang (2001), Chong e Jun (2005) e Huang e Wang (2006).

Também na seleção de variáveis em bancos de dados de espectros de infravermelho (NIR), que podem apresentar milhares de variáveis e são o tema específico deste artigo, PLS tem sido utilizada com sucesso. Hoskuldsson (2001) utiliza como critério de seleção das variáveis a sua correlação com a variável de resposta y. Variáveis apresentando baixa correlação com y são as primeiras eliminadas através de um teste de significância estatística; as variáveis restantes são inseridas em um modelo de regressão PLS. No trabalho de Xu *et al.* (2007), vetores de ponderação indicam a contribuição de cada variável, sendo utilizados nas iterações do algoritmo PLS. Vetores de ponderação são estimados por meio de uma técnica de otimização baseada em busca, designada por *Particle Swarm Optimization* (PSO).

No trabalho de Lima *et al.* (2005), variáveis irrelevantes em bancos de dados NIR são eliminadas com base na magnitude dos coeficientes da regressão PLS; o processo de remoção utiliza informações oriundas da função erro. Breiman (1995, 1996) e Kondylis e Whittaker (2010) alertam que os resultados de seleção de variáveis com abordagens baseadas estritamente na magnitude dos coeficientes de regressão PLS devem ser cuidadosamente analisados. Os autores afirmam que tais procedimentos simplistas não levam em conta a correlação entre as demais variáveis e podem resultar em modelos instáveis; tais afirmações são corroboradas por Sorol *et al.* (2010) em diversas aplicações NIR.

Uma síntese de abordagens para seleção de variáveis com fins de predição são apresentadas na Tabela 5.1. Percebe-se um predomínio de métodos baseados na regressão PLS.

Tabela 5.1: Resumo de abordagens para seleção de variáveis com fins de predição

VARIABLE SELECTION FOR PREDICTION				
PLS regression		**Genetic Algorithm**	**Stepwise**	**Clustering**
Chong and Jun [8]	Kondylis and Whittaker [23]	Gauchi and Chagnon [5]	Chong and Jun [8]	Gauchi and Chagnon [5]
Wang et al. [32]	Camacho [45]	Huang and Wang [15]	Xu and Zhang [6]	Roy and Roy [62]
Kohonen et al. [29]	Hoskuldsson [33]	Ferrand et al. [50]	Furnols et al. [53]	Hernández et al. [64]
Wold et al. [25]	Xu et al. [34]	Fei et al. [51]		
Hoskuldsson [33]	Lima et al. [46]	Leardi et al. [52]		
Xu et al. [34]	Breiman [47]	Sorol et al. [49]	**PCA**	**Lasso Regression**
Xiaobo et al. [35]	Breiman [48]	Goodarzi et al. [63]	Camacho [45]	Chong and Jun [8]
Baumann et al. [36]	Sorol et al. [49]	Felkel et al. [65]	Baumann et al. [36]	
Zhai et al. [37]	Ferrand et al. [50]	Gualdron et al. [72]	Hernández et al. [64]	
Xu et al. [38]	Fei et al. [51]	Wiegand et al. [73]		
Deeb [39]	Leardi et al. [52]			
Gauchi and Chagnon [5]	Furnols et al. [53]	**Artificial Neural Networks**	**Bayesian**	**PCR**
Xu and Zhang [6]	Gosselin et al. [54]	Fei et al. [51]	Philips and Guttman [77]	Zarzo and Ferrer [43]
Huang and Wang [15]	Zhai et al. [37]	Jiao and Li [7]	Hibbert and Armstrong [78]	
Lazraq et al. [40]	Baumann et al. [36]	Zhang [69]		
Sarabia et al. [41]	Roy and Roy [62]	Gualdron et al. [72]		
Forina et al. [42]	Hernández et al. [64]			
Zarzo and Ferrer [43]	Felkel et al. [65]			

Fonte: Anzanello e Fogliatto (2011)

O presente trabalho propõe um procedimento sequencial para a seleção de um subconjunto de variáveis no âmbito da regressão PLS, além de um critério de parada que torna possível selecionar um número apropriado de variáveis. O método utiliza como ponto de partida o CovSel, proposto por Roger *et al.* (2011). O critério de informação de Akaike, utilizado como critério de parada, avalia se a nova variável a ser introduzida é relevante na melhoria do ajuste do modelo estatístico. A abordagem proposta destaca-se pela sua simplicidade e bons resultados.

O método proposto traz duas contribuições importantes na área de seleção de variáveis para fins de predição. Primeiro, trata-se de um método concebido para seleção de variáveis em bancos de dados altamente multicolineares. A ideia é maximizar a correlação entre as variáveis selecionadas e a variável de resposta, e minimizar a correlação entre as variáveis retidas na análise. O conjunto de variáveis retidas resulta em modelos de regressão com menor variância de predição, se comparado a variáveis selecionadas exclusivamente com base em sua relação com a variável de resposta. Segundo, a proposição do critério de informação de Akaike como critério de parada no procedimento de seleção de variáveis leva a um conjunto mais eficiente de variáveis retidas, em termos de sua utilização posterior (isto é, a obtenção de um modelo de regressão vinculando variáveis retidas e variáveis de resposta).

O restante deste documento contém, na seção 5.2, a teoria sobre o método proposto. Dois bancos de dados nos quais o método é aplicado são apresentados na Seção 5.3. A seção 5.4 mostra os resultados numéricos da aplicação. A conclusão encerra o artigo na Seção 5.5.

5.2 TEORIA

5.2.1 Notação

A notação utilizada neste trabalho é descrita a seguir. Matrizes e vetores são escritos em negrito: matrizes com letras maiúsculas e vetores com letras minúsculas. Seja **X** uma matriz composta de P $(p = 1, \dots, P)$ variáveis preditoras (colunas), observadas em N $(n = 1, \dots, N)$ indivíduos (linhas), e **Y** uma matriz composta de M $(m = 1, \dots, M)$ variáveis de respostas (colunas) medida em n indivíduos (linhas).

5.2.2 Algoritmo

O algoritmo é implementado em quatro etapas, executadas iterativamente e resumidas a seguir. Inicia-se pela padronização das matrizes **X** e **Y** (exceto em casos de dados de espectros infravermelho onde a padronização não é recomendada). A variável preditora padronizada com maior covariância com **Y** e com as demais variáveis em **X** é identificada; seja X_t a representação desta variável. A informação colinear em X_t é então deflacionada de **X** e **Y** através de projeção ortogonal. Em seguida, o melhor modelo de regressão relacionando X_t às respostas em **Y** é determinado, assim como o valor do critério de informação de Akaike (AIC) correspondente. Estas operações são repetidas até que todos os p preditores, reordenados de acordo com a iteração na qual foram selecionados no algoritmo, sejam contemplados. O melhor subconjunto de variáveis a ser retido para fins de predição é finalmente determinado analisando os valores de AIC das variáveis ordenadas. Estas etapas são detalhadas a seguir.

Passo 1: Pré-tratamento das matrizes **X** *e* **Y**

Para minimizar as fontes de variação no conjunto de dados, recomenda-se padronizar as matrizes **X** e **Y**. A padronização é realizada subtraindo-se a média das linhas das matrizes de suas linhas correspondentes e dividindo as linhas resultantes por seus desvios-padrões. Este procedimento é indicado na utilização de bancos de dados de espectro infravermelho e

denominado SNV (*Standard Normal Variation*). Outra fonte de variação está associada com diferenças de escala. Para sanar este problema, fatores de escala isotrópicos são utilizados ($\propto_k$). Neste trabalho, utiliza-se o fator de escala proposto por Kunert e Qannari (1999) cuja ideia geral é a seguinte (Rossini et al, 2010). Multiplica-se o conjunto de dados **X** por $\propto_k$ para (*i*) estreitar configurações com grande amplitude de escala (para tal, utilizar $\propto_k$ < 1), ou (*ii*) expandir configurações com pequenas amplitudes de escala (para tal, utilizar $\propto_k$ > 1).

O fator de escala $\propto_k$ é assim calculado. Determinar t_k, que é a variância total do conjunto de dados **X**, adicionando as variâncias de cada coluna de **X**. Calcular t, que é a média de t_k ($k = 1, \dots, m$). O fator de escala é dado por $\propto_k = \sqrt{\frac{t}{t_k}}$, sendo específico para cada coluna (i.e., variável) da matriz de dados. Multiplicando cada coluna de **X** pelo seu respectivo fator de escala $\propto_k$, obtém-se uma nova matriz com variância total igual a t.

Passo 2: Determinar a variável mais próxima das respostas e variáveis remanescentes

A seleção das variáveis é realizada calculando o critério V_t, obtido na *t*-ésima iteração do algoritmo, e dado na equação (1).

$$V_t = Max_p \left[diag \left(\mathbf{X}^{*t} (\mathbf{X}|\mathbf{Y}) \left((\mathbf{X}|\mathbf{Y}) \right)^t \mathbf{X} \right) \right] \quad (1)$$

onde X|Y é a matriz resultante da união das matrizes X e Y, lado a lado. O critério seleciona a variável mais correlacionada tanto com as demais variáveis em X, quanto com as respostas em Y. Trata-se de uma extensão do critério CovSel. Seja $\mathrm{X_{V_t}}$ a variável a ser selecionada (com coluna $\mathrm{x_{V_t}}$ correspondente em X).

Passo 3: Deflacionar a informação de X_{V_t} das matrizes **X** *e* **Y** *através de projeção ortogonal*

O processo de deflação é equivalente a regredir as matrizes **Y** e **X** na variável X_{V_t}, de tal forma que toda informação colinear em X_{V_t} seja capturada. As colunas originais de **Y** e **X** são então substituídas pelos resíduos de seus modelos de regressão correspondentes, calculados para cada coluna de entrada. O processo é realizado pela projeção do vetor coluna $\mathbf{x}_{V_t}$ ortogonalmente em **Y** e **X**, através das seguintes operações:

$$\mathbf{P} = \mathbf{I} - \frac{\mathbf{x}_{V_t}\mathbf{x}_{V_t}^T}{\mathbf{x}_{V_t}^T\mathbf{x}_{V_t}} \quad (2)$$

$$\mathbf{X}_d = \mathbf{P}\mathbf{X} \quad (3)$$

$$\mathbf{Y}_d = \mathbf{P}\mathbf{Y} \quad (4)$$

onde $\mathbf{I}$ denota uma matriz identidade de dimensão $(n \times n)$, na eq. (2). Note que $\mathbf{x}_{V_t} = \mathbf{0}$ em $\mathbf{X}_d$.

A seguir, retorna-se ao passo *2* para determinar a próxima variável a ser selecionada. Para tanto, deve-se substituir as matrizes $\mathbf{X}$ e $\mathbf{Y}$ por $\mathbf{X}_d$ e $\mathbf{Y}_d$ na eq. (1). Os passos 2 e 3 são reiterados até que todas as P variáveis sejam selecionadas. Finalmente, reordenam-se as variáveis em $\mathbf{X}$ para refletir a iteração em que foram selecionadas.

É importante notar que usando $\mathbf{X}_d$ e $\mathbf{Y}_d$ na eq. (1), em todas as iterações após a primeira, a próxima variável selecionada será aquela que apresentar a menor correlação com a variável selecionada na iteração anterior. Assim, o primeiro conjunto de variáveis selecionadas no procedimento irá descrever as direções independentes de variância em $\mathbf{X}$, enquanto o segundo conjunto será composto de variáveis colineares ao primeiro conjunto. O passo 4 descreve uma proposta para identificação da fronteira entre esses dois conjuntos.

Passo 4: Critério de Informação de Akaike (AIC)

O Critério de Informação de Akaike – AIC (Harrell, Jr., 2001; p. 202) fornece um indicador de adequação do ajuste, utilizado na identificação do melhor modelo de regressão de um grupo de candidatos. AIC visa minimizar dois subcritérios ao selecionar o melhor modelo: (*i*) erros $\hat{\varepsilon}_n$, dados pela diferença entre os valores observados e previstos de resposta, e (*ii*) o número total de parâmetros K utilizados no modelo. O subcritério (*i*) garante um modelo bem ajustado aos dados; o subcritério (*ii*) dirige a seleção para o modelo mais parcimonioso, penalizando o excesso de ajuste (e consequente modelagem do erro experimental).

O passo atual consiste em determinar o modelo de regressão para cada iteração dos passos 2 e 3. Na iteração t, os modelos relacionam cada resposta em $\mathbf{Y}$ com o conjunto de variáveis $\{X_t, X_{t-1}, \ldots, X_1\}$; o melhor modelo é determinado através de um procedimento

stepwise. Após todas as iterações realizadas, um total de $P \times M$ modelos estarão disponíveis para análise. A seguinte estatística é calculada a cada iteração:

$$AIC_t = \sum_{m=1}^{M}\{[nlog(\sum_{n=1}^{N}\hat{\epsilon}_n^2/n)] + 2K\}, t = 1, \dots, T \quad (5)$$

Ao calcular AIC_t, todas as variáveis selecionadas em **X** até a iteração t são usadas como regressores nos M modelos obtidos naquela iteração. Assim, AIC_t fornece um critério de parada para o processo de seleção, de modo que o menor valor de AIC_t indica o conjunto de variáveis que devem ser reservadas para uso posterior. Esse conjunto de variáveis leva ao melhor modelo de previsão, consideradas todas as respostas em **Y**.

5.2.3 Método CovSel

O método utilizado para fins de comparação do método descrito na seção 5.2.2 é o proposto por Roger *et al.* (2011) denominado CovSel. CovSel é um processo iterativo de seleção de variáveis com etapas semelhantes ao método proposto diferindo somente no critério utilizado para a seleção da variável e no critério de seleção do subconjunto ótimo de variáveis.

CovSel utiliza o índice I_1, descrito na equação (6), que identifica e seleciona a variável preditora em X que melhor prediz a variável de resposta Y.

$$I_1 = ArgMax_i[\text{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X})] \quad (6)$$

A etapa seguinte à seleção corresponde a deflação das matrizes equivalente ao método proposto.

5.3 MATERIAIS E MÉTODOS

Nesta seção, o algoritmo proposto na seção 5.2.2 é aplicado a dois bancos de dados experimentais. O objetivo é verificar seu desempenho em comparação ao CovSel (Roger *et al.*, 2011). Os dois bancos de dados são compostos por leituras de espectro infravermelho. O primeiro é referente a amostras de damascos com uma única resposta; o segundo contém amostras de milho e também apresenta uma variável de resposta.

Em ambos os conjuntos o objetivo foi otimizar a predição, determinando o subconjunto de variáveis que melhor descreve a resposta. O índice V_t foi utilizado para

classificar as variáveis, enquanto que o criterio AIC_t foi utilizado para avaliar o desempenho do subconjunto de variáveis retidas a cada iteração. A validação dos resultados ocorreu através da validação cruzada *leave-one-out* (LOOCV). Este procedimento envolve a utilização de uma única observação da amostra original para a validação dos resultados. As demais observações são utilizadas como amostras para calibração. Este processo é repetido de forma iterativa até que cada observação daa amostra seja utilizada para validação e o número de repetições da LOOCV compreende 100 vezes. O processo de validação foi usado para calcular os erros padrão de validação (RMSECV - *Root-Mean-Square Error of Cross-Validation*).

- *Banco de dados de damasco*

O banco de dados experimental consiste de 731 espectros de infravermelho, obtidos de damascos em diferentes comprimentos de onda, totalizando $P = 292$ variáveis. O objetivo do experimento é determinar se os dados espectrais contêm informações úteis sobre o grau Brix, que avalia o teor de sólidos solúveis na amostra, o que foi medido em cada fruta e tratado como resposta Y. Neste banco de dados as variáveis da matriz **X** são altamente colineares: 100% dos 84.972 pares de correlações são significativas ao nível de 5% ou inferior. A correlação média é 0,8851.

- *Banco de dados de milho*

O banco de dados consiste em espectros de infravermelho de 80 amostras de milho, contendo comprimentos de onda entre 1100 e 2498 nm, com um total de $P = 700$ variáveis. O conteúdo de proteína nas amostras foi tratado como variável de resposta Y. Neste banco de dados, as variáveis da matriz **X** são ainda mais colineares do que no banco de dados anteriormente descrito: 100% dos pares de correlações são significativas ao nível de 5% ou inferior, e a correlação média é 0,9884.

5.4 RESULTADOS E DISCUSSÃO

Nesta seção são apresentados os resultados obtidos da aplicação do método na seção 5.2 (identificado como MP - método proposto) aos conjuntos de dados na seção 5.3. Para fins de comparação, foi aplicado o CovSel sobre os mesmos conjuntos de dados, sendo analisados os resultados.

As Figuras 5.1 e 5.2 mostram a evolução da variância explicada em **X** e *Y* à medida que as variáveis são selecionadas, em cada método e conjunto de dados. Conforme pode ser observado na Figura 5.1, (A) e (B), o MP obtém melhor desempenho no tocante a capturar a informação relativa à **X** e *Y* ao reter as primeiras variáveis. Este resultado pode ser explicado através do critério de seleção que cada método utiliza, uma vez que MP abrange tanto a relação entre as variáveis preditoras de **X** quanto a variável de resposta *Y*, ao passo que o critério CovSel é focado exclusivamente na relação entre **X** e *Y*. Na Fig. 5.1(A) observa-se que poucas variáveis (quatro) são suficientes para capturar a maior parte da variância (≥ 90%) de **X** no caso de MP; e é necessária a retenção de seis variáveis para capturar variância semelhante utilizando o CovSel. Os resultados da Fig. 5.1(B) indicam que é preciso reter um dez variáveis no MP para representar adequadamente a resposta *Y* e onze para CovSel.

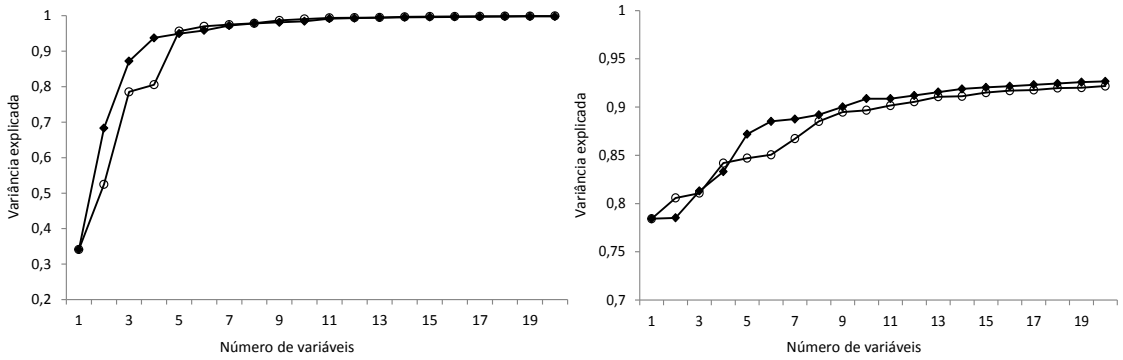


Figura 5.1: Evolução da variância explicada em **X** (A) e *Y* (B) para o banco de dados de damasco, usando (o) CovSel e (■) MP

Os gráficos da Figura 5.2 apresentam os resultados da aplicação de ambos os métodos no banco de dados de milho. À esquerda [Fig. 5.1(A)], a evolução da variância explicada em **X** é apresentada; observa-se que ambos os métodos apresentam resultados semelhantes, ainda que MP obtenha uma ótima representação dos dados com apenas duas variáveis retidas. Na Fig. 5.1(B), CovSel se sobressai ao MP após a seleção da oitava variável.

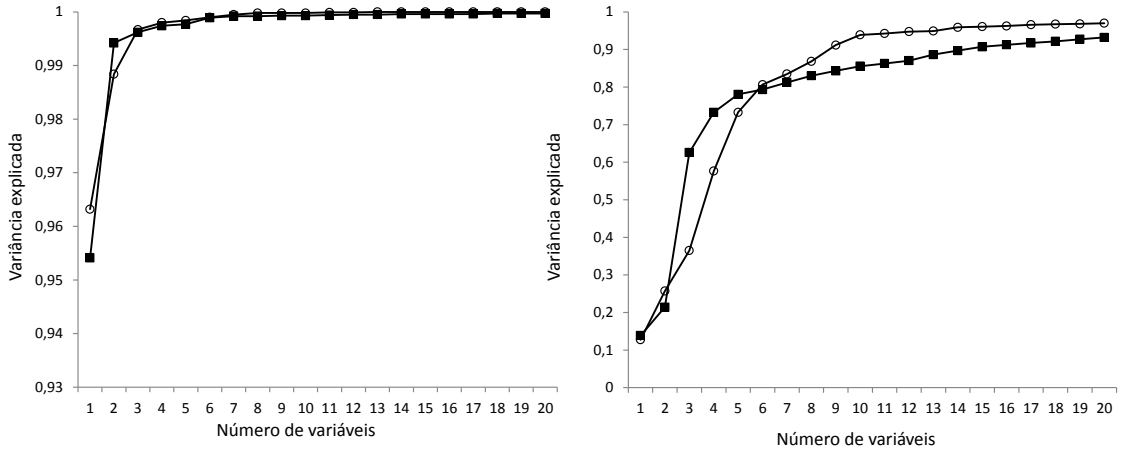


Figura 5.2: Evolução da variância explicada em **X** (A) e Y (B) para banco de dados de milho, usando (o) CovSel and (■) MP

A Figura 5.3 apresenta a evolução de três índices relacionados à adequação do ajuste dos modelos de regressão às variáveis selecionadas do banco de dados de damasco, usando CovSel e MP; são eles: coeficiente de determinação (R^2), raiz quadrada do erro médio após efetuado o processo de validação cruzada dos resultados (RMSECV), e o critério de informação de Akaike (*AIC*). *AIC* é também recomendado como critério de parada no método proposto para seleção de variáveis. Os gráficos convergem em seu comportamento: MP apresenta um desempenho geral melhor, com valores superiores nos 3 índices obtidos quando 6 variáveis são retidas; CovSel apresenta resultados inferiores, necessitando aproximadamente 17 variáveis para obter valores similares para os índices.

A Figura 5.4 apresenta a evolução dos mesmos índices descritos acima quando CovSel e PM são aplicados ao banco de dados de milho. O uso do MP produz resultados superiores em todos os casos. O índice de determinação R^2 sugere a retenção de 4 e 10 variáveis para o MP e CovSel, respectivamente ($R^2 > 0{,}9$). O número de variáveis a serem retidas analisando os dois índices seguintes são os mesmos, para os dois métodos: 8 variáveis no caso do MP e 11 variáveis utilizando CovSel. Mesmo considerando um maior número de variáveis, CovSel não atinge valores semelhantes ao MP, indicando que este possui um desempenho superior.

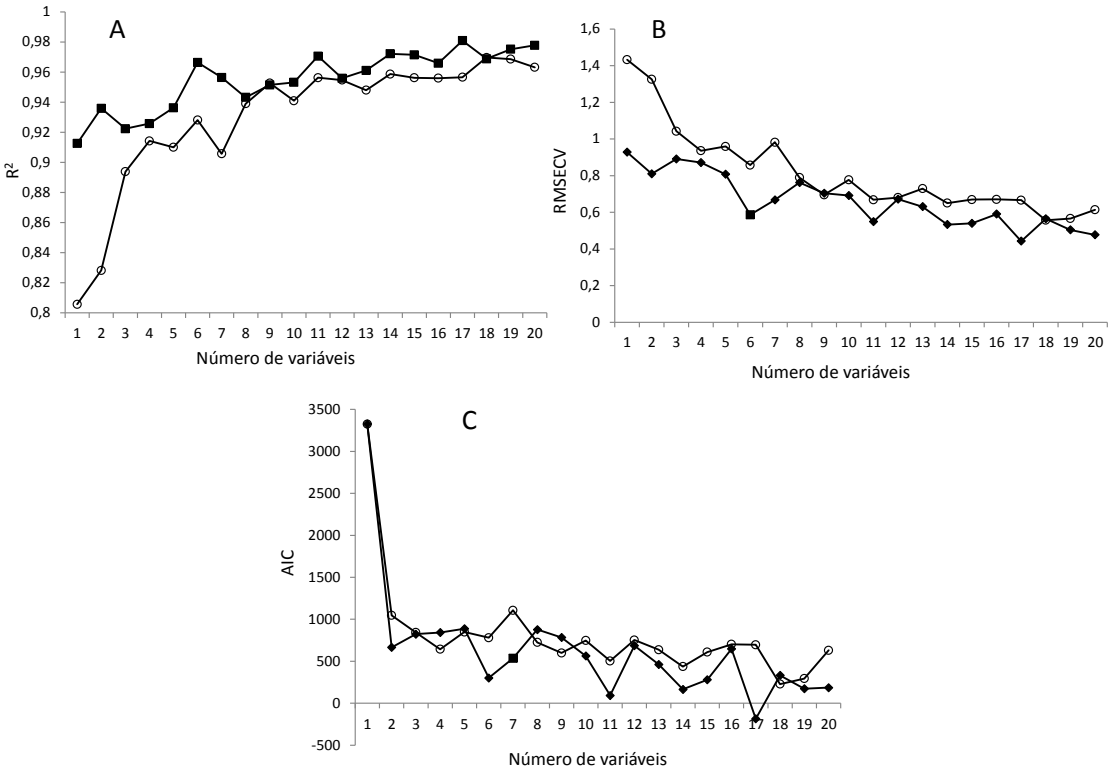


Figura 5.3: Evolução dos índices R^2 (A), RMSECV (B) e AIC (C) à medida que as variáveis são selecionadas para o banco de dados de damasco utilizando (o) CovSel e (■) MP

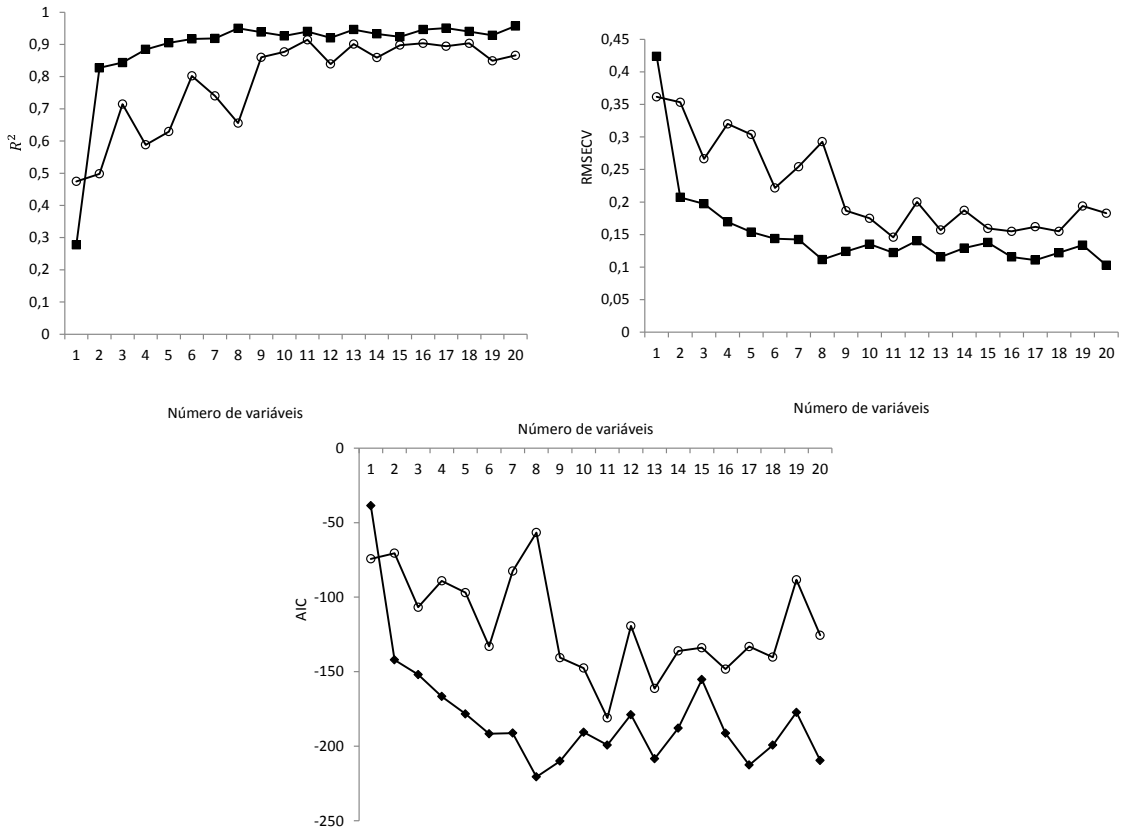


Figura 5.4: Evolução dos índices MSE (A), R^2 (B), e AIC (C) à medida que as variáveis são selecionadas para o banco de dados de milho, utilizando (o) CovSel e (■) MP

O índice AIC é o critério de parada no método proposto, conforme apresentado na seção 5.2. Analisando os gráficos das Figuras 5.3(C) e 5.4(C), alguns pontos de inflexão podem ser visualizados à medida que o índice evolui. Por exemplo, na Fig. 5.3(C) o método CovSel gera pontos de inflexão quando 4, 6, 9, 11, 14 e 18 variáveis são retidas. Tais pontos são determinados visualmente, correspondendo a situações onde o índice pára de decrescer e começa a aumentar seu valor momentaneamente.

A Tabela 5.1 lista os 6 primeiros pontos de inflexão identificados pela evolução do AIC nos dois métodos e bancos de dados. Os valores correspondentes de AIC também são exibidos. No conjunto de dados de damasco, o quarto ponto de inflexão parece ser a melhor escolha de parada em MP; já em CovSel, a melhor configuração se dá no sexto pico. Todos os resultados apresentados foram obtidos posteriormente à validação cruzada, como descrito na seção 5.2.

Tabela 5.2: Número de variáveis retidas e valores de desempenho para os primeiros pontos de inflexão do índice AIC

	MP					*CovSel*			
Dados	**N° variáveis**	**R^2**	**RMSECV**	**AIC**		**N° variáveis**	**R^2**	**RMSECV**	**AIC**
Damasco	2	0,936	0,810	**666,275**		4	0,914	0,936	**643,350**
	4	0,926	0,871	**842,173**		6	0,928	0,858	**779,988**
	6	0,966	0,587	**301,891**		9	0,953	0,695	**598,510**
	11	0,971	0,549	**93,221**		11	0,956	0,669	**504,011**
	14	0,972	0,533	**165,171**		14	0,959	0,650	**436,737**
	17	0,981	0,443	**-185,178**		18	0,970	0,557	**227,914**
Milho	6	0,931	0,144	**-191,614**		3	0,655	0,266	**-106,764**
	8	0,891	0,112	**-220,584**		6	0,746	0,222	**-133,043**
	11	0,932	0,122	**-199,237**		11	0,907	0,146	**-180,981**
	13	0,938	0,116	**-208,430**		13	0,873	0,157	**-161,269**
	17	0,927	0,111	**-212,630**		16	0,878	0,155	**-148,29**
	-	-	-	**-**		18	0,8846	0,155	**-140,266**

Dado que o critério de parada é regido pelo índice AIC, o MP, quando aplicado ao banco de dados de damasco, apresenta seu mínimo quando 17 variáveis são retidas (-185,178), sendo este também o ponto que gera o menor RMSECV (0,443). No entanto, a situação na qual 11 variáveis são selecionadas apresenta uma configuração satisfatória quando comparado ao método CovSel, uma vez que este gera resultados inferiores nos 3 índices avaliados, em qualquer dos pontos de inflexão.

Resultado semelhante é encontrado para o banco de dados de milho. O MP atinge sua configuração ótima quando seleciona 8 variáveis (pico 2), enquanto a melhor configuração de CovSel se apresenta no terceiro pico (11 variáveis). É interessante ressaltar que os resultados do MP são superiores aos de CovSel, independente do pico utilizado.

5.5 CONCLUSÃO

A seleção de variáveis é um tema presente em análises químicas devido às inúmeras variáveis coletadas. Este artigo propôs um método para seleção de variáveis em bancos de dados com alta multicolinearidade, contendo uma variável de resposta, com fins de predição. O objetivo foi selecionar um subconjunto de variáveis com a maior covariância tanto com **Y** (matriz das variáveis de resposta) quanto com as demais variáveis em **X** (variáveis de predição). Também foi proposta a utilização do critério de informação de Akaike como critério de parada no procedimento de seleção. O método proposto foi comparado com o

CovSel de Roger *et al.* (2011), e aplicado em dois bancos de dados de espectro infravermelho, obtidos de amostras de damasco e milho.

Os resultados obtidos permitiram a redução de 292 para 17 variáveis no caso do banco de dados de damasco, e de 700 para 8, no caso do banco de dados de milho. Além disso, os resultados obtidos para o método proposto mostraram-se superiores aos apresentados pelo CovSel, em todas as instâncias de comparação. Futuras pesquisas incluem o desenvolvimento de alternativas para selecionar o melhor conjunto de variáveis para predição com múltiplas variáveis de respostas.

5.6 REFERÊNCIAS

ALMOY, T., A simulation study on comparison of prediction methods when only a few components are relevant. **Computational Statistics & Data Analysis**, v.21, p.87-107, 1996.

ANZANELLO, M. J, FOGLIATTO, F. S, ROSSINI, K., Data mining-based method for identifying discriminant atributes in sensory profiling. **Food Quality and Preference**, v.22, p.139-148, 2011.

BAUMANN K., ALBERTO, H., KORFF, M., A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part I. Search algorithm, theory and simulations. **Journal of Chemometrics**, v.16, p.339-350, 2000.

BREIMAN, L., Better subset selection using the nonnegative garrote. **Technometrics**, v.37, p.373–384, 1995.

BREIMAN, L., Heuristics of instability and stabilization in model selection, **Annals of Statistics**, v.24, p.2350–2383, 1996.

CAMACHO, J., Missing-data theory in the context of exploratory data analysis. **Chemometrics and Intelligent Laboratory Systems**, v.103, p.8-18, 2010.

CHONG, I., JUN, C., Performance of some variable selection methods when multicollinearity is present. **Chemometrics and Intelligent Laboratory Systems**, v.78, p.103-112, 2005.

DEEB, O., Correlation ranking and stepwise regression procedures in principal components artificial neural networks modeling with application to predict toxic activity and human serum albumin binding affinity. **Chemometrics and Intelligent Laboratory Systems**, v.104, p.181-194, 2010.

DINGSTAD, G., WESTAD, F., NAES,T., Three case studies illustrating the properties of ordinary and partial least squares regression in different mixture models. **Chemometrics and Intelligent Laboratory Systems**, v.71, p. 33-45, 2004.

FELKEL, Y., DÖRR, N., GLATZ, F., VARMUZA, K., Determination of the total acid number (TAN) of used gas engine oils by IR and chemometrics applying a combined strategy for variable selection. **Chemometrics and Intelligent Laboratory Systems**, v.101, p.14-22, 2010.

FERRAND, M., HUQUET, B., BARBEY, S., BARILLET, F., FAUCON, F., LARROQUE, H., LERAY, O., TROMMENSCHLAGER, J.M., BROCHARD, M., Determination of fatty acid profile in cow's milk using mid-infrared spectrometry: Interest of applying a variable selection by genetic algorithms before a PLS regression. **Chemometrics and Intelligent Laboratory Systems**, 2010. In press.

FORINA, M., CASOLINO, C., MILLAN, C., Iterative predictor weighting (IPW) PLS: a technique for the elimination of useless predictors in regression problems. **Chemometrics and Intelligent Laboratory Systems**, v.13, p.165-184, 1999.

FURNOLS, M., TERAN, M., GISPERT, M., Estimation of lean meat content in pig carcasses using X-ray Computed Tomography and PLS regression. **Chemometrics and Intelligent Laboratory Systems**, v.98, p.31-37, 2009.

GAUCHI, J., CHAGNON, P., Comparison of selection methods of exploratory variables in PLS regression with application to manufacturing process data, **Chemometrics and Intelligent Laboratory Systems**, v.58, p.171-193, 2001.

GOODARZI, M., FREITAS, M., WU, C. H., DUCHOWICZ, P., pKa modeling and prediction of a series of pH indicators through genetic algorithm-least square support vector regression. **Chemometrics and Intelligent Laboratory Systems**, v.101, p.102-109, 2010.

GOSSELIN, R., RODRIGUE, D., DUCHESNE, C., A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications. **Chemometrics and Intelligent Laboratory Systems**, v.100, p.12-21, 2010.

GUALDRON, O., LIOBET, E., BREZMES, J., VILANOVA, X., CORREIG, X., Fast variable selection for gas sensing applications. **Proceedings IEEE Sensors**, p.892-895, 2004.

HARRELL, JR.F.E., **Regression modeling strategies with applications to linear models, logistic regression, and survival analysis**. Springer-Verlag: New York, 568p. (2001).

HERNÁNDEZ, N., KIRALJ, R., FERREIRA, M., TALAVERA, I., Critical comparative analysis, validation and interpretation of SVM and PLS regression models in a QSAR study on HIV-1 protease inhibitors. **Chemometrics and Intelligent Laboratory Systems**, v.98, p.65-77, 2009.

HIBBERT, D., ARMSTRONG, N., An introduction to Bayesian methods for analyzing chemistry data: Part II: A review of applications of Bayesian methods in chemistry. **Chemometrics and Intelligent Laboratory Systems**, v.97, p.211-220, 2009.

HOSKULDSSON, A., Variable and subset selection in PLS regression. **Chemometrics and Intelligent Laboratory Systems**, v.55, p. 23-38, 2001.

HUANG, C., WANG, C., GA-based feature selection and parameters optimization for support vector machines. **Expert Systems Applications**, v.31, p.231-240, 2006

JIAO, L., LI, H., QSPR studies on the aqueous solubility of PCDD/Fs by using artificial neural network combined with stepwise regression. **Chemometrics and Intelligent Laboratory Systems**, v.103, p.90-95, 2010.

KOHONEN, J., REINIKAINEN,S., AALJOKI, K., HÖSKULDSSON, A., Non-linear PLS approach in score surface. **Chemometrics and Intelligent Laboratory Systems**, v.97, p.159-163, 2009.

Kondylis, A., Whittaker, J., Adaptively preconditioned Krylov spaces to identify irrelevant predictors. **Chemometrics and Intelligent Laboratory Systems**, v.104, p.205-213, 2010.

LAZRAQ, A., CLEROUX, R., GAUCHI, J., Selecting both latent and exploratory variables in the PLS1 regression model. **Chemometrics and Intelligent Laboratory Systems**, v.66, p.117-126, 2003.

LEARDI, R., SEASHOLTZ, M., PELL, R., Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data. **Analytica Chimica Acta**, v.461, p.189-200, 2002.

LEE, H., LEE, M., PARK, J., Multi-scale extension of PLS algorithm for advanced on-line process monitoring. **Chemometrics and Intelligent Laboratory Systems**, v.98, p.201-212, 2009.

LIMA, S., MELLO, C., POPPI, R., PLS pruning: a new approach to variable selection for multivariate calibration based on Hessian matrix of errors. **Chemometrics and Intelligent Laboratory Systems**, v.76, p.73-78, 2005.

MUTEKI, K., MACGREGOR, J., Multi-block PLS modeling for L-shape data structures with applications to mixture modeling. **Chemometrics and Intelligent Laboratory Systems**, v.85, p.186-194, 2007.

PHILIPS, R., GUTTMAN, I., A new criterion for variable selection. **Statistics & Probability Letters**, v.38, p.11-19, 1998.

Q. Fei, M. Li, B. Wang, Y. Huan, G.Feng, Y. Ren, Analysis of cefalexin with NIR spectrometry coupled to artificial neural networks with modified genetic algorithm for wavelength selection. **Chemometrics and Intelligent Laboratory Systems**, v.97, p.127-131, 2009.

ROGER, J. M., PALAGOS, B., BERTRAND, D., FERNANDEZ-AHUMADA, E., CovSel: Varible selection for highly multivariate and multi-response calibration Application to IR spectroscopy. **Chemometrics and Intelligent Laboratory Systems**, v.106, p.216-223, 2011.

ROY, P., ROY, K., On Some Aspects of Variable Selection for Partial Least Squares Regression Models. **QSAR Combinatorial Science**, v.27, p.302-313, 2008.

SAHMER, K., QANNARI, E.M., Procedures for the selection of a subset of attributes in sensory profiling. **Food Quality and Preference**, v.19, p.141-145, 2008.

SARABIA, L., ORTIZ, M., SANCHEZ, A., Dimension wise selection in partial least squares regression a bootstrap estimated signal-noise relation to weight the loadings, in: PLS and Related Methods, **Proc. PLS'01 International Symposium**, CISIA-CERESTA Editeur, Paris, p.327-339, 2001.

SOROL, N., ARANCIBIA, E., BORTOLATO, S., OLIVIERI, A., Visible/near infrared-partial least-squares analysis of Brix in sugar cane juice: A test field for variable selection methods. **Chemometrics and Intelligent Laboratory Systems**, v.102, p.100-109, 2010.

WANG, D., SRINIVASAN, R., LIU, J., P. GURU, LEONG, K., Data-driven soft sensor approach for quality prediction in a refinery process, **Proceeding IEEE International Conference Ind. Inf**. 2006.

WIEGAND, P., PELL, R., COMAS, E., Simultaneous variable selection and outlier detection using a robust genetic algorithm. **Chemometrics and Intelligent Laboratory Systems**, v.98, p.108-114, 2009.

WOLD, S., SJOSTROM, M., ERIKSSON, L., L. PLS-regression: a basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v.58, p.109-130, 2001a.

WOLD, S., TRYGG, J., BERGLUND, A. H., Antti, Some recent developments in PLS modeling. **Chemometrics and Intelligent Laboratory Systems**, v.58, p.131-150, 2001b.

WU, W., MANNE, R., Fast regression methods in a Lanczos (or PLS-1) basis. Theory and applications. **Chemometrics and Intelligent Laboratory Systems**, v.51, p.145-161, 2000.

XIAOBO, Z., JIEWEN, Z., HOLMES, M., HANPIN, M., JIYONG, S., XIAOPIN, Y., YANXIAO, L., Independent component analysis in information extraction from visible/near-

infrared hyperspectral imaging data of cucumber leaves. **Chemometrics and Intelligent Laboratory Systems**, v.104, p.265-270, 2010.

XU, J., LIANG, H., CHEN, B., XU, W., SHEN, X., LIU, H., Linear and nonlinear QSPR models to predict refractive indices of polymers from cyclic dimer structures. **Chemometrics and Intelligent Laboratory Systems**, v.92, p.152-156, 2008.

XU, L., JIANG, J., WU, H., SHEN, G., YU, R., Variable-weighted PLS. **Chemometrics and Intelligent Laboratory Systems**, v.85, p.140-143, 2007.

XU, L., ZHANG, W., Comparison of different methods for variable selection. **Analytica Chimica Acta**, v.446, p.477-483, 2001.

ZARZO, M., FERRER, A., Batch process diagnosis: PLS with variable selection versus block-wise PCR. **Chemometrics and Intelligent Laboratory Systems**, v.73, p.15-27, 2004.

ZHAI, H., CHEN, X., HU, A., A new approach for the identification of important variables. **Chemometrics and Intelligent Laboratory Systems**, v.80, p.130-135, 2006.

ZHANG, Y., An improved QSPR study of standard formation enthalpies of acyclic alkanes based on artificial neural networks and genetic algorithm. **Chemometrics and Intelligent Laboratory Systems**, v.98, p.62-172, 2009.

6 ARTIGO 5 – Comparação de diferentes abordagens na avaliação sensorial e desenvolvimento de produtos alimentícios

Karina Rossini

Programa de Pós Graduação em Engenharia de Produção – UFRGS
Av. Osvaldo Aranha, 99 – 5° andar, 90035-190 - Porto Alegre– RS
e-mail: karinarossini@hotmail.com

Flávio Sanson Fogliatto

Programa de Pós Graduação em Engenharia de Produção – UFRGS
Av. Osvaldo Aranha, 99 – 5° andar, 90035-190 - Porto Alegre– RS
e-mail: ffogliatto@producao.ufrgs.br

Resumo

Este trabalho apresenta um comparativo entre dois métodos usados na análise sensorial de formulações de um tipo de produto achocolatado em pó. É feita a comparação entre um método tradicional, o Teste de Ordenação Individual, e uma abordagem qualitativa baseada em Grupos Focados. A forma como estes métodos foram implementados é descrita detalhadamente. Cinco formulações foram avaliadas por ambos os métodos, que revelaram concordância em seus resultados. Embora os métodos tenham conduzido a conclusões similares, o uso do Grupo Focado se sobressai pela riqueza das informações complementares que oferece.

Palavras chave: Análise sensorial, grupos focados, achocolatado em pó.

6.1 INTRODUÇÃO

A crescente globalização da economia e o aumento da diversidade e da variedade de produtos faz com que o consumidor se torne mais seletivo, exigindo melhor qualidade dos

produtos manufaturados. As empresas, para permanecerem sustentáveis, precisam inovar e desenvolver produtos que antecipem as necessidades dos consumidores, ganhando mercado frente à concorrência.

Nas empresas do setor de alimentos, a avaliação sensorial apresenta-se como um componente crítico para o processo de desenvolvimento de novos produtos (SIDEL; STONE, 1993). A análise sensorial utiliza princípios provenientes da ciência de alimentos, fisiologia, psicologia e estatística, fornecendo respostas objetivas para as propriedades de alimentos e de como são percebidas pelos cinco sentidos: visão, olfato, tato, audição e paladar (PIGGOTT; SIMPSON; WILLIAMS, 1998).

Há uma grande variedade de métodos disponíveis para as avaliações sensoriais, desde simples, como, por exemplo, testes clássicos de discriminação, até complexos, como teste de tempo-intensidade. Tais métodos são aplicados desde o desenvolvimento até diferentes estágios do ciclo de vida de um produto (KOEFERLI; SCHWEGLER; CHEN, 1998). Além disso, a avaliação sensorial pode ser utilizada para verificar alterações ocorridas durante o período de vida de prateleira, efeitos da embalagem, variações de matérias-primas e qualidade no processamento dos produtos (MURRAY; DELAHUNTY; BAXTER, 2001).

Diante do exposto, justifica-se o interesse na investigação de diferentes métodos e sua adequação a cenários específicos. Assim, este artigo apresenta uma contribuição relevante: a comparação de dois métodos para coleta e análise de dados sensoriais com vistas a dar suporte ao processo de desenvolvimento de um novo produto aplicados a empresas de alimentos. Os dois métodos utilizados diferem pela natureza de suas abordagens: uma tradicional, individual, representada pelo Teste de Ordenação de Preferência; outra tipicamente qualitativa, onde se aplica o método de Grupo Focado. Embora inúmeros relatos sobre análise sensorial estejam descritos na literatura, o estudo aqui proposto inexistente na literatura até então.

O presente trabalho é composto de cinco seções. Além da introdução, é apresentado um referencial teórico, seguido pela descrição dos métodos empregados para realização da coleta, tratamento e análise dos dados e, posteriormente, apresenta-se o estudo aplicado, finalizando com as conclusões do mesmo.

6.2 REFERENCIAL TEÓRICO

Em pesquisas sensoriais e de consumo um painel de avaliadores é frequentemente utilizado para estudar propriedades de certos produtos, como por exemplo, produtos alimentícios (Dijksterhuis, 1995). A avaliação sensorial envolve um conjunto de técnicas que medem atributos sensoriais a partir de respostas humanas, conforme percebidas pelos sentidos humanos. As informações obtidas através das avaliações sensoriais podem ser utilizadas pelas empresas como suporte técnico para pesquisa, industrialização, marketing e controle de qualidade, para sustentar decisões administrativas, diminuindo o risco que acompanha o processo de tomada de decisão. Relativo ao consumidor, a análise sensorial assegura que produtos industriais cheguem ao mercado com um conceito suficientemente desenvolvido, como atributos sensoriais que atendam suas expectativas. Complementarmente, a avaliação sensorial fornece diretrizes para preparar e servir amostras em condições controladas, de acordo com o objetivo da avaliação e o tipo de produto a ser avaliado, minimizando fontes de variação (LAWELESS e HEYMANN, 1998).

A avaliação sensorial é efetuada de maneira científica quando se utiliza um painel sensorial composto por um grupo de pessoas (degustadores, julgadores ou avaliadores), selecionadas para analisar os atributos sensoriais dos alimentos e treinadas de acordo com o objetivo da avaliação e o tipo de produto a ser avaliado (MELIGAARD; CIVILLE e CARR, 1999).

Os métodos sensoriais podem ser divididos em métodos discriminativos, descritivos e subjetivos ou afetivos (STONE E SIDEL, 1993). Os testes discriminativos devem ser utilizados com o objetivo de determinar se duas amostras são perceptivelmente diferentes, pois, conforme Laweless e Heymann (1998), dois produtos podem ser compostos de formulações diferentes, mas sem perceptível diferença aos consumidores.

Os métodos descritivos estão entre as ferramentas mais elaboradas da ciência sensorial e envolvem a detecção (discriminação) e descrição tanto de componentes sensoriais qualitativos quanto quantitativos por um painel treinado de julgadores. Os aspectos qualitativos dos produtos incluem aroma, aparência, sabor, textura e as propriedades sonoras dos produtos. Os julgadores sensoriais devem quantificar os aspectos dos produtos de maneira a facilitar a descrição da percepção dos seus atributos. As análises sensoriais descritivas podem ser utilizadas, dentre outras aplicações, no controle de qualidade e na comparação de

protótipos de produtos para compreender as respostas dos consumidores em relação aos seus atributos sensoriais. A principal vantagem da análise descritiva está em sua habilidade de permitir que seja determinada uma relação entre medidas sensoriais descritivas e a instrumental ou de preferência do consumidor (MURRAY *et al.*, 2001).

Os métodos subjetivos ou afetivos medem o quanto uma população gostou de um produto. Eles são utilizados para avaliar sua preferência ou aceitabilidade. Preferência pode ser definida como a expressão do grau de gostar ou a escolha de uma amostra em relação à outra. Aceitabilidade, por sua vez, pode ser descrita como uma experiência caracterizada por uma atitude positiva, e pela utilização atual do produto, isto é, o hábito de comprar ou consumir um alimento (FARIA; YOTSUYANAGY, 2002).

Os testes afetivos podem ser classificados em testes qualitativos e quantitativos. Os testes qualitativos são aqueles que avaliam subjetivamente as respostas de uma amostra de consumidores em relação às propriedades sensoriais de um produto, expectativas relacionadas à embalagem ou propaganda, etc., ou simplesmente na investigação detalhada de seus hábitos, atitudes e expectativas em relação a um tema ou produto alimentício. Consistem de entrevistas em profundidade, em geral com até 50 consumidores, ou em grupo, através de grupo focado (DUTCOSKY, 2007).

Conforme Ribeiro (2007), a técnica de grupos focados é mais bem empregada para geração de ideias e impressões que se tem de um produto ou serviço do que para examiná-los sistematicamente. Assim sendo, essa técnica não deveria ser utilizada para determinar a proporção de pessoas que pensam de determinado modo ou de outro. Os resultados não são representativos da população, por isso não são projetáveis, dificultando a codificação, tabulação e a análise estatística. Meligaard *et al.* (1999) afirmam que para o grupo focado são selecionados de 10 a 12 consumidores com base em critério específico de interesse, como sequência de uso do produto, idade, etc. A reunião, com duração de 1 a 2 horas, é conduzida pelo moderador do grupo, o qual apresenta o tópico de interesse e facilita a discussão usando técnicas de dinâmica de grupo, visando conseguir o máximo possível de informações específicas dos participantes sobre o objeto da reunião.

Os métodos qualitativos são aplicados quando se busca um posicionamento inicial do consumidor em relação ao conceito de um produto ou de um protótipo. Na fase de desenvolvimento de conceito e planejamento, grupos focados são utilizados para obter

informações provenientes da interação individual com interesse ou prática comum (DUTCOSKY, 2007).

Os testes quantitativos são aqueles que avaliam a resposta de um grande grupo de consumidores a uma série de perguntas que visam determinar o grau de aceitabilidade global de um produto, identificar fatores sensoriais que determinam a preferência ou medir respostas específicas a atributos sensoriais de um produto (DUTCOSKY, 2007).

Os principais métodos subjetivos referem-se aos testes de ordenação, pareado ou ordenação múltiplas, aceitabilidade e escala hedônica.

6.3 METODOLOGIA

No presente estudo, para fins comparativos, foram utilizadas dois métodos de análise sensorial: o Teste Individual de Ordenação de Preferência, realizado individualmente, e a discussão em Grupos Focados.

Passo 1: Determinar a equipe multidisciplinar

Para coleta de dados e aplicação dos testes acima citados, utilizou-se uma equipe multidisciplinar composta por dez pessoas, de ambos os sexos e com faixa etária entre 20 e 35 anos, com conhecimentos prévios sobre análise sensorial e consumidores regulares de achocolatado em pó. A equipe foi dividida em dois subgrupos (A e B), com cinco pessoas cada, a fim de facilitar as discussões dos grupos focados bem como possibilitar a alternância da aplicação dos métodos.

Passo 2: Coleta de dados

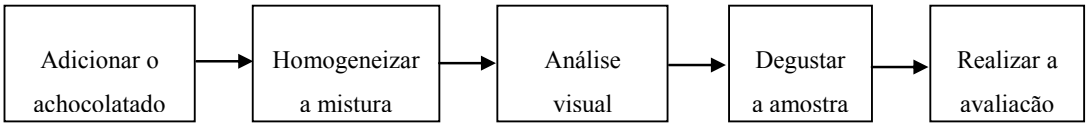
A coleta de dados foi realizada em dois momentos, com um intervalo de sete dias entre eles. Na primeira etapa o subgrupo A participou da reunião do grupo focado e o subgrupo B realizou a análise sensorial individual. Na segunda etapa, o subgrupo A realizou a avaliação sensorial individual e o subgrupo B participou da reunião de grupo focado. A etapa de coleta individual foi efetuada em duplicata.

Passo 3: Preparação das amostras e para a análise sensorial

O conjunto de amostras foi composto de duas marcas líderes de mercado e três formulações de teste da empresa. O preparo das mesmas ocorreu previamente à aplicação das técnicas e obedeceu às instruções quanto ao modo de preparo sugerido pelo respectivo fabricante.

Os participantes receberam as cinco amostras codificadas com algarismos aleatórios (904, 238, 385, 162 e 691). Cada amostra foi composta por uma medida de leite e outra de pó dispostos de forma separada, permitindo a visualização e análise, por parte dos degustadores, tanto das características do pó quanto da mistura.

Para a análise sensorial, cada participante seguiu uma sequência pré-determinada, repetindo-a para as cinco amostras, como descrito a seguir:



Observando o procedimento padrão para esta análise, os provadores realizaram a análise sensorial das amostras da esquerda para a direita e bebendo água entre as degustações. Este procedimento foi realizado tanto no grupo focado quanto na avaliação individual.

Passo 4: Grupos focados

As reuniões de grupo focado foram realizadas na sala de reuniões da empresa. Os participantes ficaram dispostos envoltos ao redor de uma mesa redonda, diante das amostras. Utilizou-se um *flip-chart* para anotações de informações e ordenação final de preferência das amostras. As reuniões foram registradas, também, através de gravador de áudio.

As reuniões foram conduzidas e coordenadas por um moderador, o qual iniciou as atividades apresentando-se, expondo os objetivos do encontro, a que se destinavam as informações coletadas e salientando a importância da colaboração e transparência do grupo. Posteriormente, realizou-se uma breve explanação sobre a técnica de grupos focados e análise sensorial, ressaltando o procedimento adequado para a realização da mesma. Após a orientação inicial, os participantes realizaram a análise sensorial, conforme descrito no Passo

3, seguida da discussão do grupo, culminando com a ordenação consensual de preferência das amostras. Cada grupo focado teve duração média de 1 hora.

Passo 5: Teste Individual de Ordenação de Preferência

A análise individual compreendeu o preenchimento da ficha de avaliação e uma ordenação de preferência das amostras. O procedimento de preparo das amostras bem como análise e degustação foi semelhante ao utilizado durante o grupo focado.

Passo 6: Análise de dados

Os dados analisados compreenderam as informações levantadas durante o grupo focado, ou seja, as ordenações finais realizadas pelos subgrupos e as percepções discutidas, bem como as ordenações individuais. A partir das ordenações individuais gerou-se uma tabela de pontuação a qual possibilitou uma análise estatística através teste de Friedman, o qual utiliza a tabela de Newel e MacFarlane com nível de significância estatística de 5% (DUTCOSKY, S., 2007).

6.4 ESTUDO APLICADO

6.4.1 Descrição dos dados coletados

6.4.1.1. Descrição dos grupos focados

A descrição dos dados coletados compete, primeiramente, às reuniões dos grupos focados. Nestes, após a explicação dos objetivos e do método do trabalho, questões introdutórias foram formuladas. Estas versaram sobre o conceito de achocolatado em pó para os participantes, as principais características avaliadas pelo consumir e os fatores que influenciam a compra de um achocolatado em pó.

O primeiro grupo focado iniciou a discussão enfatizando as alterações no paladar desde o início do consumo (ainda quando crianças) do achocolatado, onde a intensidade do sabor adocicado era o fator predominante na decisão de compra. Informaram, também, as marcas as quais consumem fielmente (sendo citados o Nescau e o Toddy Light). Entre os aspectos importantes e decisivos na compra do produto estão: (*i*) a forte intensidade de cor, assemelhando-se a chocolate e não café com leite; (*ii*) o gosto e aroma de chocolate, e não baunilha ou caramelo; (*iii*) o gosto de chocolate se sobrepondo ao sabor doce; e (*iv*) pó que

forneça cremosidade ao leite e que tenha capacidade de solubilizar-se mesmo quando adicionado ao leite frio, embora a permanência de "bolinhas" de pó não dissolvidas no leite não fosse considerada problema.

Após a discussão da questão introdutória, iniciou-se a análise sensorial das amostras de achocolatado em pó conforme procedimento relatado na metodologia. Alguns participantes relataram o consumo do pó "puro", então, nestes casos, eles degustaram, também, o pó antes de adicioná-lo. De forma paralela à degustação, procederam-se os comentários descritos na Tabela 6.1.

Tabela 6.1: Descrição das amostras segundo o primeiro grupo focado

Amostra	Características	Comentários
Amostra 904	- intensidade da cor ótima, semelhante ao chocolate - aroma de caramelo -"sabor doce" adequado - boa solubilização do pó - sabor residual desagradável no pó	- "...na verdade, esperamos que o leite adquira de maneira instantânea a coloração e, em alguns casos, é necessário a adição de muito pó para que o leite lembre chocolate, frustrando as expectativas do consumidor, no caso desta amostra, a cor foi bastante satisfatória".
Amostra 238	- ausência de aroma no pó do achocolatado - solubilização instantânea	-"... a cor do leite ficou parecendo café com leite e não chocolate, o sabor lembra caramelo e não possui qualquer aroma".
Amostra 385	- características semelhantes à amostra 238, exceto nos quesitos doçura, no qual esta se mostrou mais doce, e tonalidade mais clara que a 238	

Amostra 162	- tonalidade clara do pó - boa dissolução do pó - o pó conferiu cremosidade ao leite, sabor e aroma de chocolate mais acentuado que as amostras 238 e 385, porém menos que a 904 - embora a cor do pó fosse fraca, conferiu uma cor agradável ao leite, lembrando chocolate	-"... textura aveludada interessante e granulometria do pó menor, não conseguindo diferenciar o açúcar".
Amostra 691	- pó muito doce, aparecendo o açúcar e sem qualquer aroma predominante	-"... muito doce, sem gosto de chocolate; resumindo, não é bom".

De forma geral, os degustadores concluíram que as amostras 238, 385 e 691 apresentaram-se muito semelhantes, onde o gosto do leite estava predominante ao de chocolate praticamente inexistia. A amostra 904 era a que impunha mais dificuldade de dissolução, mas continha maior sabor de chocolate e tonalidade mais escura. A 162 proporcionava um pó mais "encorpado", o qual era conferido devido à menor granulometria e maior uniformidade do pó.

Após término das observações procedeu-se a ordenação consensual das amostras, tal que a preferência decrescia com a ordenação. Cabe salientar que os avaliadores foram quase unânimes na seleção das melhores amostras; porém, a classificação das demais foi objeto de debate, devido à semelhança observada entre elas. Obteve-se a seguinte ordenação final: 1º – 904; 2º – 162; 3º – 691; 4º – 385; 5º – 238.

O procedimento efetuado com o segundo grupo focado foi idêntico ao primeiro. Após a apresentação dos objetivos e observações iniciais, o grupo iniciou a discussão relatando os aspectos importantes analisados ao consumir achocolatado, tais como: "ser escuro como Toddinho, dissolver bem no leite gelado, sabor de chocolate, não ser tão doce (mas também que não precise adicionar açúcar) e aroma de chocolate e não baunilha, doce de

leite ou caramelo". Informaram, também, as marcas consumidas atualmente, sendo citado principalmente o Nescau.

Após a discussão da questão introdutória, iniciou-se a degustação das amostras conforme orientação já descrita na metodologia. As amostras foram dispostas na mesma ordem do Grupo Focado anterior. Apesar dos códigos das amostras terem sido trocados no segundo grupo, a descrição dos resultados segue a mesma codificação do primeiro grupo focado, conforme apresentado na Tabela 6.2.

Tabela 6.2: Descrição das amostras de acordo com o segundo grupo focado

Amostra	Características	Comentários
Amostra 904	- cor ótima, muito próxima do ideal - aroma e gosto de chocolate - pouca solubilidade do pó	- "... restaram algumas "bolinhas" na superfície e o leite não adquiriu corpo".
Amostra 238	- identificação dos grânulos de açúcar - cor mais clara do pó - sabor exageradamente doce	- "...só possui cheiro de leite". -"...pó desceu, a cor é de café com leite e é muito doce".
Amostra 385	- identificação das partículas de açúcar no pó do achocolatado - produto com boa solubilização - tonalidade mais clara que a amostra anterior - semelhança com a amostra 238, com preferência desta a aquela	- "...gosto do sabor". - "...doçura e sabor bom".
Amostra 162	- cremosidade na mistura - aroma agradável	-"... cor de achocolatado para bolo".

	- solubilidade intermediária	-"... pela cor do pó não compraria". - "...gosto não é tão bom, muito pesado".
Amostra 691	- aroma de baunilha - cor semelhante a 162	- "...cheiro de leite". - "...leite com muito açúcar". -"...sem gosto de chocolate, produto aguado".

Semelhante ao primeiro grupo focado, os degustadores concluíram que as amostras 238, 385 e 691 apresentaram-se muito semelhantes. Ao realizar a ordenação das amostras, este grupo apresentou-se mais homogêneo do que o anterior; a ordem de preferência das amostras foi: 1º – 904; 2º – 162; 3º – 385; 4º – 238; 5º – 691.

6.4.1.2. Descrição do teste de ordenação individual

Além dos grupos focados realizou-se a avaliação individual das amostras através do teste subjetivo de ordenação de preferência, como descrito anteriormente. Esta análise individual foi realizada em duplicata, de modo que cada participante realizou a análise duas vezes com um intervalo entre elas. Em quaisquer das situações, grupos focados e análises individuais, a sequência das amostras era a mesma, porém codificadas com algarismos diferentes. A Tabela 6.3 retrata as ordenações individuais. Os valores de 1 a 10 referem-se aos julgadores.

Tabela 6.3: Ordenação individual, em duplicata, de cada degustador

Participante/ Ordenação	1		2		3		4		5		6		7		8		9		10	
1º	904	904	162	162	162	162	904	904	904	904	904	691	904	238	904	904	904	904	904	904
2º	162	162	904	904	904	904	162	385	162	691	691	162	691	904	162	162	162	162	162	238
3º	385	691	385	385	691	691	385	238	238	385	162	904	162	162	385	691	385	385	691	385
4º	238	385	691	238	385	385	691	691	385	162	238	238	385	691	238	238	238	691	238	691
5º	691	238	238	691	238	238	238	162	691	238	385	385	238	385	691	385	691	162	385	162

6.4.2 Análise e Interpretação dos dados coletados

6.4.2.1. Resultados dos grupos focados

Nesta etapa é apresentada a análise dos resultados anteriormente descritos. As sessões de grupo focado foram unificadas e as discussões reestruturadas em uma tabela comparativa (Tabela 6.4).

Tabela 6.4: Comparação entre as percepções dos degustadores para cada amostra analisada durante os grupos focados

	Pó			Mistura (pó+leite)				
Característica/Amostra	Cor	Homegeneidade	Aroma	Solubilidade	Cor mistura	Gosto	Doçura	Consistência
904	Adequada, ótima	Boa	Caramelo, bom	Boa, algumas bolinhas na superfície, um pouco a desejar	Ótima, semelhante ao chocolate, instantânea	Chocolate	Bom	Pequena
162	Muito clara, não compraria	Muito boa, aveludada	Chocolate	Boa	Boa	Chocolate (mais que 385, 691 e 238 e menos que 904)	Bom	Boa
691	Semelhante a 238	Identificação de grânulos de açúcar	Ausente, baunilha	Boa	Fraca	Sem gosto característico (somente de leite)	Muito doce	Pequena
385	Mais clara que a 238	Identificação de grânulos de açúcar	Melhor que a 238	Semelhante a 238	Semelhante a 238, decepcionante	Semelhante a 238	Mais doce que a 238	Pequena
238	Mais clara que a 904	Identificação de grânulos de açúcar	Ausente, de leite	Boa, instantanea	Café com leite	Acentuado de leite, não lembra chocolate, leite com açúcar	Muito doce	Pequena

Através das exposições dos grupos focados, foi possível perceber que a importância das características do produto final (mistura do achocolatado em pó com o leite) era substancialmente maior do que aquela do pó na opinião dos degustadores. Por conta desta verificação, determinou-se uma relação de importância entre as características do pó e da mistura. Além disso, o desenvolvimento da tabela comparativa (Tabela 6.4) possibilitou a atribuição de notas para todas as amostras e para cada atributo avaliado. As notas tanto para as amostras quanto o peso para cada atributo variaram em uma escala de 0 a 10. O resultado desta distribuição de pesos pode ser conferido na Tabela 6.5.

Tabela 6.5: Tabela de pesos e notas para as características que compõem as amostras, de acordo com os resultados dos grupos focados

	Pó			Mistura (pó+leite)				
Peso	5	3	2	3	9	6	4	2
Característica/ Amostra	Cor	Homegeneidade	Aroma	Solubilidade	Cor mistura	Gosto	Doçura	Consistência
904	10	8	8	8	10	10	8	4
162	2	10	10	8	8	8	8	8
691	6	4	2	8	2	2	4	4
385	4	4	4	8	2	4	2	4
238	6	4	2	8	2	2	4	4

A multiplicação das avaliações realizadas nos grupos focados por seus pesos permitiu obter notas finais de cada amostra, as quais representam aproximadamente a percepção verbalizada no estudo em grupos focados. Os resultados são apresentados na Tabela 6.6.

Tabela 6.6: Avaliação das formulações obtida durante as sessões dos grupos focados

Amostra	Avaliação
904	8,94
162	7,41
691	4,18
385	4,12
238	4,18

Analisando a Tabela 6.6, percebe-se que as amostras 691, 385 e 238 atingem valores semelhantes. Comparando–se estes com os das demais amostras verifica-se diferença substancial com relação às amostras 904 e 162. Entre as amostras 904 e 162 também se observa diferença importante na avaliação final.

Comparando o resultado quantitativo, baseado nas discussões dos grupos focados, com as ordenações de preferência destes mesmos grupos, verifica-se concordância em parte dos resultados, uma vez que a ordenação dos primeiros lugares coincide em ambos os grupos focados. As ordenações das demais formulações dos grupos se confundem, informação evidenciada também na quantificação da Tabela 6.6, através da pequena diferença nos resultados.

Sendo assim, através dos grupos focados a amostra 904 é aquela que melhor atende as demandas dos consumidores, seguida pela amostra 162. As demais amostras assemelham-se entre si e revelam baixa qualidade (isto é, não atenderiam satisfatoriamente as demandas dos consumidores).

6.4.2.2. Resultados dos testes de ordenação individual

A interpretação das ordenações individuais resultantes do teste de preferência deu-se por meio do procedimento indicado pelo teste de Friedman. A escala que compõe o teste é apresentada na Tabela 6.7.

Tabela 6.7: Escala utilizada pelo teste de Friedman para análise das ordenações individuais

Ordenação	Avaliação
1º	5
2º	4
3º	3
4º	2
5º	1

Uma vez aplicada a Tabela 6.7 nas ordenações individuais, obteve-se os resultados descritos na Tabela 6.8.

Tabela 6.8: Resultado do teste de Friedman para ordenação individual das amostras

Posição/Amostra	904	162	385	238	691
1	70	20	0	5	5
2	20	36	4	4	12
3	3	9	27	6	15
4	0	2	10	16	12
5	0	3	5	8	5
Total	**93**	**70**	**46**	**39**	**49**

A partir dos totais obtidos na Tabela 6.8, realiza-se a comparação entre eles, conforme previsto no teste, a fim de verificar se existe diferença significativa entre as amostras. Assim, é possível constatar quais amostras diferem significativamente entre si, ao nível 5% de significância. Para que as amostras sejam diferentes significativamente, a diferença entre seus totais deve ser superior a 28. A Tabela 6.9 explicita a comparção acima mencionada.

Tabela 6.9: Resultado da comparação das diferenças entre os totais de cada amostra

Amostras	904	162	385	238	691
904	0				
162	23	0			
385	**47**	24	0		
238	**54**	**31**	7	0	
691	**44**	21	3	10	0

Concluiu-se que o resultado da análise individual converge com o apresentado pelos grupos focados: a amostra 904 não difere significativamente da amostra 162, porém difere significativamente das demais (691, 385 e 238). A amostra 162 difere significativamente da amostra 238. As amostras 691, 385 e 238, por sua vez, não diferem entre si. Assim, a amostra que, ao ser consumida, melhor atende as demandas do consumidor é a 904, seguida pela 162. As demais amostras não apresentam diferença estatística significativa, além de apresentarem escores baixos de preferência.

6.5 CONCLUSÕES

Este trabalho teve como proposta apresentar um estudo comparativo entre dois métodos diferentes de análise sensorial na área de alimentos. As abordagens analisadas, ambas de cunho qualitativo, baseiam-se em princípios diferentes. O Teste de Ordenação de Preferência, o primeiro dos métodos analisados, é aplicado de forma individual; nos Grupos Focados, em contrapartida, ocorre uma discussão em grupo sobre as amostras que estão sendo avaliadas.

O uso de ambos os métodos conduziu ao mesmo resultado, identificando uma formulação superior, que atendia as demandas dos usuários e uma segunda formulação com avaliação intermediária. As demais formulações resultaram inferiores, não diferiam significativamente entre si e não atendiam as demandas do consumidor.

Embora se tenha observado similaridade nos resultados gerados com ambos os métodos, o grupo focado se sobressai pela riqueza das informações propiciadas pelas discussões em grupo, as quais servem como um importante subsídio para possíveis adequações de um determinado produto. Enquanto sugestão para a área de análise sensorial, indiferentemente do método a ser utilizado na avaliação, recomenda-se que: (*i*) o mesmo seja precedido de uma discussão (em grupo) dos elementos considerados importantes a serem atendidos e avaliados; e (*ii*) após a avaliação individual, novamente seja conduzida a discussão em grupo para esclarecer os motivos das pontuações atribuídas, no caso de discrepância entre os valores obtidos individualmente e em grupo. A documentação das percepções, conforme observado neste estudo aplicado, pode constituir importante subsídio para as equipes de desenvolvimento de produto.

Agradecimentos

A autora expressa seus agradecimentos a todos que contribuíram na realização deste trabalho em especial à empresa que permitiu o desenvolvimento do mesmo.

6.6 REFERÊNCIAS

DIJKSTERHUIS, G. Assessing panel consonance. **Food Quality and Preference**, v.6, p.7-14, 1995.

DUTCOSKY, S. D. **Análise Sensorial de Alimentos.** 2ª Edição. Curitiba: Editora Champagnat - Coleção Exatas 4, 2007. 239 p.

DUTCOSKY, S. D. **Análise Sensorial de Alimentos.** 2ª Edição. Curitiba: Editora Champagnat - Coleção Exatas 4, 2007. 239 p.

FARIA, E. V.; YOTSUYANAGY, K. **Técnicas de Análise Sensorial**. 1ª Edição. Campinas: ITAL/LAFISE, 2002. 116p.

KOEFERLI, C. S.; SCHWEGLER, P. P.; CHEN, D. H. Application of classical and novel sensory techniques in product optimization. **Lebensm.-Wiss. u.-Technology**, v.31, p.407-417, 1998.

LAWELESS, H. T.; HEYMANN, H. **Sensory Evaluation of Food: Principles and Practices**. New York: Chapman & Hall, 1998. 819p.

MELIGAARD, M.; CIVILLE, G. V.; CARR, B. T. **Sensory Evaluation Techniques**. 3ª Edição. Boca Raton: CRC Press, 1999. 387p.

MURRAY, J. M.; DELAHUNTY, C. M.; BAXTER, I. A. Descriptive sensory analysis: past, present and future. **Food Research International**, v.34, p.461-471, 2001.

PIGGOTT, J. R.; SIMPSON, S. J.; WILLIAMS, S. A. R. Sensory analysis. **International Journal of Food Science and Technology**, v.33, p.7-18, 1998.

RIBEIRO, J. L. D. **Grupos Focados: teoria e aplicações**. 2ª Edição. Porto Alegre: FEENG/UFRGS, 2007. 93p.

SIDEL, L. J.; STONE, H. The role of sensory evaluation in the food industry. **Food Quality and Preference**, v.4, p. 65-73, 1993.

7 CONSIDERAÇÕES FINAIS

Este capítulo apresenta as conclusões da tese, além de sugestões para trabalhos futuros.

7.1 CONCLUSÕES

A presente tese teve por objetivo o desenvolvimento de metodologias de seleção de variáveis que contribuam na análise de dados oriundos de avaliações sensoriais descritivas e de espectro infravermelho, nas indústrias de alimentos e química.

O primeiro objetivo específico declarado, **estudar as principais técnicas de análise multivariada de dados, como são comumente organizadas e como podem contribuir no processo de seleção de variáveis**, é descrito nos referenciais teóricos dos quatro primeiros artigos apresentados. Cada artigo versou sobre as técnicas multivariadas especificas e pertinentes ao seu tema.

Em relação aos objetivos específicos: **identificar e estruturar técnicas de análise multivariada de dados de forma a construir um modelo que reduza o número de variáveis necessárias para fins de caracterização, classificação e predição dos produtos; reduzir a lista de variáveis/atributos, selecionando aqueles relevantes e não redundantes, reduzindo o tempo de execução e a fadiga imposta aos membros de um painel em avaliações sensoriais e análises quimiométricas; e validar o modelo proposto utilizando dados reais,** os artigos 1, 2, 3 e 4 propõem o uso combinado de diferentes técnicas multivariadas, de forma a atingi-los.

O artigo 1 prevê a junção da Análise de Componentes Principais (PCA) com a Análise Discriminante (AD) em processo iterativo para eliminação dos atributos menos relevantes em estudo de analise sensorial descritiva. A PCA foi utilizada para identificar atributos importantes através dos pesos por ela gerados, e a AD foi utilizada na classificação das amostras. O número de atributos a serem retidos foi determinado de forma subjetiva em um gráfico de acurácia. O método foi avaliado em um estudo de caso, mostrando-se adequado, uma vez que conduziu a uma redução significativa no número de atributos retidos, com acurácia similar à máxima possível, obtida mediante retenção da totalidade dos atributos.

O artigo 2 faz uso do mesmo banco de dados do artigo 1 e possui objetivo semelhante, exceto por contemplar simultaneamente a seleção de julgadores e de atributos. No artigo, a proposta foi utilizar métodos de projeção multivariada (PCA), além de ferramentas de mineração de dados (KNN) de forma iterativa. A seleção de julgadores se deu através de um índice de consistência proposto por Ledauphin *et al*. (2006). O método é recomendado para situações em que se deseja selecionar tanto julgadores quanto atributos.

O artigo 3, que objetivou discriminar produtos e julgadores, propõe o uso da ferramenta Análise Discriminante em Mínimos Quadrados Parciais (PLS-DA). O artigo realizou uma comparação entre quatro métodos de análise [PLS-DA, PCA com o banco de dados completo, PCA com valores de médias, e Análise Canônica (CVA)], observando a estabilidade dos mesmos frente à reamostragem dos julgadores. O método propõe aplicar PLS-DA no banco de dados, retendo r componentes; em seguida, uma reamostragem dos julgadores é realizada e o PLS-DA é reaplicado. Após repetir diversas vezes esses passos, elipses de confiança são geradas possibilitando verificar a estabilidade do método na seleção de variáveis. Posterior à etapa de comparação de métodos, o artigo propôs o uso do índice gerado pelo PLS-DA, o *Variable Importance in the Projection* (VIP) como forma de priorizar os atributos, por apresentar resultados mais estáveis que os demais métodos.

O artigo 4 segue o raciocínio do artigo 3 no sentido de explorar a ferramenta PLS. No entanto, neste artigo o PLS serve de sustentação teórica uma vez que o critério para seleção de variáveis é baseada na máxima covariância tanto entre as variáveis independentes (preditoras) quanto com a variável dependente (de resposta), para dados altamente multicolineares e um número muito grande de variáveis. O método foi comparado com o CovSel, de Roger et al (2011), o qual leva em consideração somente a correlação entre as variáveis dependentes e as variáveis de resposta. A seleção de variáveis foi feita com fins de predição e o estudo de caso ocorreu em dois bancos de dados de espectro de infravermelho. A totalidade das correlações apresentadas por estes bancos de dados é significativa. O método apresentou-se coerente e adequado para dados multicolineares, selecionando um menor número de variáveis e apresentando índices de adequação de ajuste dos modelos aos dados superiores em relação ao CovSel.

Em relação ao objetivo específico **comparar diferentes abordagens de análise sensorial voltadas ao desenvolvimento de novos produtos**, o artigo 5 comparou dois métodos para coleta e análise de dados sensoriais que diferiam pela natureza de suas

abordagens: uma individual tradicional (Teste de Ordenação de Preferência); outra tipicamente qualitativa (Grupos Focados). A comparação dos testes é relevante, pois, permite identificar a superioridade dos Grupos Focados através da riqueza das informações por eles fornecidas.

7.2 SUGESTÕES PARA TRABALHOS FUTUROS

Pesquisas futuras podem ser desenvolvidas como extensões dos desenvolvimentos aqui propostos. São elas:

a) Propor um modelo de seleção de variáveis para bancos de dados com múltiplas respostas (múltiplas variáveis dependentes);
b) Propor novos índices de importância para atributos e julgadores;
c) Propor novos critérios de parada para retenção de atributos;
d) Aplicar o modelo de seleção de variáveis baseado na máxima covariância a outros bancos de dados, com características distintas; e
e) Estender o uso dos métodos de seleção de variáveis para variáveis de processos.

7.3 REFERÊNCIAS

LEDAUPHIN, S.; HANAFI, M.; QANNARI, E.M. Assessment of the agreement among the subjects in fixed vocabulary profiling. **Food Quality and Preference**, v.17, p.277-280, 2006.

ROGER, J.M., PALAGOS, B., BERTRAND, D., FERNANDEZ-AHUMADA, E.CovSel: Varible selection for highly multivariate and multi-response calibration Application to IR spectroscopy, **Chemometrics and Intelligent Laboratory Systems**, v.106, p.216-223, 2011.

8 Apêndice

Versão em inglês do artigo 2 – “Método baseado na mineração de dados para identificação de atributos discriminantes em perfis sensoriais”.

Food Quality and Preference 22 (2011) 139–148

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Data mining-based method for identifying discriminant attributes in sensory profiling

Michel J. Anzanello, Flavio S. Fogliatto *, Karina Rossini

Federal Univ of Rio Grande do Sul, Industrial Engineering, Av Osvaldo Aranha, 99 – 5o andar, 90035-190 Porto Alegre, RS, Brazil

ARTICLE INFO

Article history:
Received 20 November 2009
Received in revised form 24 August 2010
Accepted 26 August 2010
Available online 6 September 2010

Keywords:
Attribute selection
Discriminant attributes
Sample classification
Data mining tools

ABSTRACT

Selection of attributes from a group of candidates to be assessed through sensory analysis is an important issue when planning sensory panels. In attribute selection it is desirable to reduce the list of those to be presented to panelists to avoid fatigue, minimize costs and save time. In some applications the goal is to keep attributes that are relevant and non-redundant in the sensory characterization of products. In this paper, however, we are interested in keeping attributes that best discriminate between products. For that we present a data mining-based method for attribute selection in descriptive sensory panels, such as those used in the Quantitative Descriptive Analysis. The proposed method is implemented using Principal Component Analysis and the *k*-Nearest Neighbor classification technique, in conjunction with Pareto Optimal analysis. Objectives are (*i*) to identity the set of attributes that best discriminate samples analyzed in the panel, and (*ii*) to indicate the group of panelists that provide consistent evaluations. The method is illustrated through a case study where beef cubes in stew, used as combat ration by the American Army, are characterized in sensory panels using the Spectrum protocol.

1. Introduction

Descriptive Analysis (DA) methods aim at providing sensory profiles of products. In essence, DA protocols call for the evaluation of samples regarding the intensity of a typically large set of attributes. Samples are evaluated individually, and results are expressed on a continuous numerical scale. There is a single number indicating the attribute intensity for each sample, and sensory panel data can thus be treated as quantitative data.

Although widely used by practitioners (Murray, Delahunty, & Baxter, 2001), DA methods present some drawbacks. First, the number of attributes to be assessed by panelists is usually large; some DA profiling protocols may present up to 30 attributes (Carbonell, Izquierdo, & Carbonell, 2007). As a result, data collection tends to be tiring, time consuming and costly. Second, it is not guaranteed that a thorough profiling of samples will include attributes with respect to which samples may be discriminated, although in some applications that is the main purpose of the data collection; e.g. Granitto, Gasperi, Biasioli, Trainotti, and Furlanello (2007). Third, DA methods do not offer any structured way to score panelists.

Attribute selection is an important research topic in the area of sensory evaluation. Selection may be aimed at (*i*) identifying a subset of non-redundant attributes that best describe products, or at (*ii*) finding attributes that best discriminate between products. With respect to objective (*i*), authors such as Dijksterhuis, Frost, and Byrne (2002), Westad, Hersleth, Lea, and Martens (2003), and Sahmer and Qannari (2008) have proposed the use of several multivariate projection methods to select a subset of relevant and/or non-redundant attributes from a larger group. With respect to objective (*ii*), and methodologically aligned with the propositions in this paper, Granitto et al. (2007) introduced the use of data mining tools (more specifically, Random forests heuristics) to select attributes that best discriminate products. In all cases, sensory profiling was used to characterize products and the drawbacks listed above were partially addressed.

In this paper we propose the combined use of multivariate projection methods and data mining tools to select relevant attributes in sensory profiling. We consider relevant attributes those that best discriminate products evaluated in a panel. Our method is implemented in six steps. First, panelists are ranked using a consistency index, and those whose evaluations differ from the rest of the group have their data omitted from the dataset, using a "leave one panelist out at a time" approach. *S* datasets are produced, each comprised of evaluations from a subgroup of panelists; the minimum number of panelists in a subgroup is user-defined. Remaining steps are implemented *S* times, one for each dataset obtained in the initial step. For a given dataset we run a Principal Component Analysis (PCA) and compute attribute importance indices based on PCA weights. We then classify products in the dataset using the

* Corresponding author. Tel.: +55 51 3308 4294; fax: +55 51 3308 4007.
E-mail address: ffogliatto@producao.ufrgs.br (F.S. Fogliatto).

0950-3293/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.foodqual.2010.08.010

k-Nearest Neighbor (KNN) algorithm and compute the classification accuracy. The attribute with the lowest importance index is removed from the dataset; products are classified again and a new accuracy value is produced. The process is repeated until there is only one attribute remaining. Accuracy results are plotted in a graph and the entire procedure is repeated for the next dataset. Results from each iteration are collapsed in a single accuracy plot, and the best subset of attributes and panelists is determined using Pareto Optimal analysis.

We envision three relevant contributions in the method we propose, as explained next.

First, we present an attribute selection method that simultaneously identifies discriminant attributes and consistent panelists. More specifically the discriminative power of different subsets of attributes is measured in the light of evaluations performed by different groups of panelists. The objective is to simultaneously identify the best set of discriminant attributes keeping those panelists whose evaluations improve the attributes' capability to discriminate between products. Other attribute selection propositions in the literature focus exclusively on the minimization of retained attributes; e.g. Granitto et al. (2007).

Second, we combine Principal Component Analysis with the KNN classification algorithm to obtain an efficient attribute selection method. Our method uses KNN as the classification technique because of its good performance in practical applications in the data mining field, conceptual simplicity, and wide availability in computational packages; see Chaovalitwongse, Fan, and Sachdeo (2007) and Anzanello, Albin, and Chaovalitwongse (2009).

Third, we use Pareto Optimal (PO) analysis to identify a limited number of distinctive solutions that maximize the classification accuracy and minimize the number of retained attributes, as well as the number of panelists. PO has been employed in a large variety of applications, such as the analysis of life cycles of chemical products in Azapagic (1999), scheduling of manufacturing operations in Taboada and Coit (2008), and reliability optimization in power transmission in Taboada and Coit (2007).

To illustrate the proposed attribute selection method, we apply it in a case study involving the military product beef stew MREs, or meals-ready-to-eat (Fogliatto, Albin, & Tepper, 1999). The product is beef stew in plastic pouches used as field rations for soldiers. The product is made at a pilot plant, funded by the DLA/DOD, and located at Rutgers University, USA. Eight different products were evaluated by a 9 member sensory panel with respect to 26 sensory attributes. Our method reduced the number of panelists and attributes needed to discriminate between products to 5 and 17, respectively.

The rest of this paper is organized as follows. Section 2 gives a brief literature review on attribute selection approaches proposed in the sensory field, and introduces some of the tools applied in the method we propose. Section 3 presents the sequence of steps for applying the method. Section 4 shows the numerical results from application of the method to the case, followed by conclusions in Section 5.

2. Background

Variable and feature selection is an important research topic in areas where large datasets are common, such as Genetics and Linguistics. The objectives of variable selection are to improve the performance of predictors, and to better describe the process generating the data. Strategies for variable selection have been reviewed by Guyon and Elisseeff (2003), with a special emphasis on ranking methods. Such methods are based on ordering variables according to an importance index, and then reducing the dataset by discarding the ones with smallest index scores (Gauchi & Chagnon, 2001).

In sensory analysis ranking is commonly performed through ANOVA using the attributes' *F*-values as an importance index. If a multivariate index is desired, projection methods such as principal component analysis (PCA) or STATIS may be used. We now review some works on attribute selection for discriminating purposes in sensory analysis. With the exception of Granitto et al. (2007) where data mining tools are applied, all approaches are based on combinations of multivariate techniques and ANOVA.

In Linear Discriminant Analysis (LDA) data are assumed to follow a multivariate normal distribution, with a common covariance matrix for all categories (Ripley, 1996). The Mahalanobis distances of each object from the centroids of the categories are computed, and objects are assigned to the category with smallest distance. The delimiter between two categories is a linear function which can be a straight line in the case of two independent variables, for example. Ranking of attributes in LDA is typically performed using the prediction error rate in a validation sample as the importance index. A traditional application of LDA in attribute selection may be found in Rason, Marin, Dufour, and Lebecque (2007), while Granitto, Biasioli, Endrizzi, and Gasperi (2008) extend the use of LDA to more elaborate methods. In comparison with our method, LDA-based attribute selection procedures do not test different subsets of discriminant attributes in search of the best subset. The procedure is to score attributes based on their performance in a validation sample and choose a cut-off point to determine the subset of attributes to be retained. On the other hand, our method iteratively assesses the quality of classification by extensively testing different subsets of attributes, in search of the best classification accuracy.

ANOVA may be used solely to determine if significant differences exist between products or panelists. However, applying ANOVA to each sensory attribute increases the probability of Type I error (Tabachnick & Fidell, 1996). To overcome that, ANOVA may be used in conjunction with PCA for attribute selection purposes. A more usual approach consists of applying ANOVA on scores obtained from a PCA on raw sensory data matrix, as reported by Chabanet (2000) and Westad et al. (2003). The data matrix in this case presents combinations of products and panelists in the rows, and attribute assessments in the columns. In case significant effects are present, Fisher LSD *post hoc* tests may be employed to explore the differences between individual products or panelists. Luciano and Naes (2009) refer to this approach as PCA-ANOVA.

Alternatively ANOVA may be used on each attribute data matrix separately, followed by PCA on the matrices of estimated main effects and interactions. Such procedure, known as ASCA (Jansen et al., 2005) uses PCA to interpret the results from the ANOVA. ASCA and the method in which PCA is followed by ANOVA are compared by Luciano and Naes (2009) in a dataset from a sensory analysis of a candy product, with similar results. In opposition to our proposed method, the aforementioned approaches neither test the performance of subsets of attributes regarding their classification power nor jointly verify the best subset of panelists.

Granitto et al. (2007) proposed the use of random forests (RFs) to select discriminating attributes in samples of cheese. RFs are sets of heuristics-created decision trees such that differences between trees are maximized. The discriminating function derived from RFs assigns importance weights to attributes; such weights may be used to select a reduced number of attributes, however keeping the discriminative power of the function at a desired level. The authors were the first to propose the use of data mining tools in the context of sensory profiling; however, in their ap-

proach panelist selection is not performed simultaneously with attribute selection, as proposed in our paper. In addition the RF technique, a rather complex classification tool if compared to the KNN algorithm proposed here, is not available in most statistical packages.

A comprehensive theoretical survey of variable selection approaches is reported in Liu and Yu (2005), while a comparison of algorithms for that purpose is presented by Kudo and Sklansky (2000): the best method combines leave-one-out rule and KNN classification technique, backing our proposition.

Some classical approaches for variable selection arise from the food manufacturing industry, although not directly dealing with sensory attribute selection. Several approaches use PCA to identify the variables with most variance and then apply tools for classification and clustering (e.g. data mining, discriminant and clustering techniques) using the selected variables; see Mallet, De Vel, and Coomans (1998) and Guo, Wu, Massart, Boucon, and Jong (2002). A method to reduce dimensionality in industrial wine manufacturing was reported by Urtubia, Perrez-Correa, Soto, and Pszczolkowski (2007): PCA was first used to select the variables carrying most of the metabolite interaction information, and classes of similar behavior were then generated by applying the *K*-means clustering technique on the lower-dimensioned dataset. In Camara, Alves, and Marques (2006), PCA was associated with discriminant analysis to differentiate and classify wines; coefficients of discriminant functions were used to identify the relevant variables.

A similar study was conducted by Rebolo, Pena, Latorre, Botana, and Herrero (2000) for testing the authenticity of wines produced in a specific region. Variables describing chemical substances were initially selected using cluster analysis and PCA techniques, followed by a Stepwise Bayesian analysis. The study also aimed at categorizing the wines in classes according to their origin. With identical purposes, Marini, Bucci, Magri, and Magri (2006) studied two classification techniques, namely Soft Independent Modeling of Class Analogies (SIMCA) and UNEQ to verify the authenticity of Italian wines. SIMCA describes the similarities among products in a category using a principal component analysis (Wold & Sjostrom, 1977), while the UNEQ is a multivariate normal class model assuming an individual dispersion (i.e. unequal dispersion) of each class, similar to a Quadratic Discriminant Analysis (Derde & Massart, 1986). The identification of the relevant variables was performed by means of a Stepwise Linear Discriminant Analysis. In Capron, Smeyers-Verbeke, and Massart (2007), a wine dataset consisting of 63 process variables was evaluated using decision trees and modifications of Partial Least Squares (PLS) regression to identify the most important variables aimed at classifying wines into 4 different classes (i.e. countries of origin).

Next we provide some insight on two of the analytical tools used in our method: KNN classification algorithm and Pareto Optimal analysis.

The *k*-Nearest Neighbor is a data mining technique for classifying objects based on closest training examples in the variable space. KNN is among the simplest algorithms for classification of observations (Duda, Hart, & Stork, 2001).

Consider observations in a *J*-dimensional dataset, corresponding to the *J* attributes, and two classes of products (A or B). The objective is to classify a new observation in A or B based only on attributes. We consider the *k*-nearest neighbors of the new observation where distance is measured by Euclidean distance. For each of the *k* neighbors identify the class, A or B. One way to classify the new observation is by majority voting: the new observation is in class A if the majority of its *k*-nearest neighbors is in A. If $k = 1$, then the observation is simply assigned to the class of its nearest neighbor. The number of neighbors, *k* (*k* is a positive integer, typically small), is selected by maximizing accuracy of classification in the dataset where the class of each observation is known. Further details about KNN classification technique can be found in Wu et al. (2008).

The advantage of the KNN is that it is conceptually simpler, more intuitive than other classification techniques, and widely available in software packages. Further, KNN requires only one parameter, *k*, and the classification accuracy is not too sensitive to this choice within a reasonable range. Due to its simplicity, KKN has been applied in a wide variety of contexts including text recognition patterns in Weiss et al. (1999), detection of abnormal brain activity in Chaovalitwongse et al. (2007), and production batches classification in Anzanello et al. (2009).

The Pareto Optimal (PO) analysis identifies a set of distinctive solutions in applications with multiple objective functions. These functions frequently do not present a unique solution, but a set of suitable solutions. Variable selection applications where classification performance measures are maximized and the number of retained variables is minimized are examples of scenarios with several possible solutions. Due to its practical applicability, the PO analysis has been widely integrated to algorithms and approaches for optimization purposes, as reported in Deb, Pratap, Agarwal and Meyarivan (2002), and Deb, Thiele, Laumanns and Zitzler (2002).

Solutions identified by the Pareto Optimal are named non-dominated, meaning that they cannot be surpassed by other neighbor solutions in the evaluated objectives. That enables significant reduction on the number of potential solutions in that the analysis can be focused on a small set of effectively better solutions. These solutions are typically illustrated on a boundary named Pareto frontier. The identification of the best solution may depend on subjective information, and may become complex as the number of objective functions increases; see Horn, Nafpliotis, and Goldberg (1994), Zitzler and Thiele (1999), and Taboada and Coit (2008) for details.

3. Method

The method to select attributes that best discriminate products relies on six operational steps: 1. Measure panelists' consistency using a suitable index. 2. Apply a multivariate technique on the dataset consisting of sensory attributes. In our method a Principal Component Analysis (PCA) is used, but other techniques may be considered. 3. Compute a vector of attribute importance indices based on PCA weights. 4. Classify the sensory dataset using the *k*-Nearest Neighbor (KNN) technique and compute the classification accuracy. Then eliminate the attribute with the lowest importance index, classify the dataset again, and re-compute the accuracy. Continue such iterative process until there is only one attribute remaining. 5. Construct an accuracy graph. 6. Remove the less consistent panelist and perform a new attribute selection, repeating steps 2 to 6. These operational steps are detailed in subsections to follow; Matlab codes used to perform analyses are given in the Appendix.

Let p ($p = 1, \ldots, P$) denote the panelists, j ($j = 1, \ldots, J$) the attributes, and i($i = 1, \ldots, I$) the products analyzed in a sensory panel. Panelists' evaluations on a product regarding all attributes may be repeated d ($d = 1, \ldots, D$) times. Consider a sensory data matrix $\mathbf{X}$ comprised of $(P \times I \times D)$ rows and J columns, with element $x_{pid,j}$ Matrix $\mathbf{X}$ may be deployed into P individual matrices $\mathbf{X}_p$, each containing assessments of a given panelist on all attributes and products. The method's objective is to correctly classify the $(P \times I \times D)$ observations in $\mathbf{X}$ into I product classes by properly choosing discriminant attributes and consistent panelists.

Step 1: Measure panelists' consistency

Measure panelists' consistency using a suitable index. We suggest using the index associated with the weighted average configuration proposed by Ledauphin, Hanafi, and Qannari (2006). The index has several interesting characteristics. First, location and dispersion effects in panelists assessments are removed through pre-treatment of the data matrix. Second, using an analytical framework similar to the STATIS method (Lavit, Escoufier, Sabatier, & Traissac, 1994) the index highlights panelists presenting a different understanding of the attributes from the rest of the group; such non-performance indication is aligned with one of our method's objectives which is to identify the group of best performing panelists. Third, the index implementation requires straightforward calculations, as presented next.

Start by centering and reducing data matrices $\mathbf{X}_p$. For that, first subtract from all column entries the column average to obtain centered data matrices $\mathbf{Xc}_p$. Next, obtain matrices $\mathbf{Y}_p$ by multiplying each matrix $\mathbf{Xc}_p$ by a scalar $\theta_p = 1/\sqrt{t_p}$, where t_p is the sum of the squares of all entries in $\mathbf{Xc}_p$ i.e. $t_p = \text{trace}(\mathbf{Xc}_p^t)\mathbf{Xc}_p$, with $\mathbf{Xc}_p^t$. denoting the transpose of $\mathbf{Xc}_p$. Formally, $\mathbf{Y}_p = \theta_p \mathbf{Xc}_p$.

The weighted average configuration of the sensory dataset is obtained as follows. Consider matrices $\mathbf{Y}_k$ and $\mathbf{Y}_l$ from panelists k and l. Determine a $(P \times P)$ matrix $\mathbf{S}$ with entries corresponding to a similarity measure between panelists k and l given by $s_{kl} = (1 + t_{kl})/2$, where $t_{kl} = \text{trace}(\mathbf{Y}_k^t \mathbf{Y}_l)$, for k, l = 1, ... ,P. Determine the eigenvector corresponding to the largest eigenvalue of $\mathbf{S}$; i.e. $\beta^t = [\beta_1, \ldots, \beta_p]$, such that $\sum_{p=1}^{p} \beta_p = 1$.

The weighted average configuration is a compromise matrix $\mathbf{C}$ that takes into account the panelists' performances. Formally, $\mathbf{C} = \sum_{p=1}^{p} \beta_p \mathbf{Y}_p$. The performance index for panelist p is given by

$$\alpha_p = \frac{\text{trace}(\mathbf{Y}_p^t \mathbf{C})}{\sqrt{trace(\mathbf{C}^t \mathbf{C})}}. \quad (1)$$

The alpha values vary in the interval from −1 to +1. A panelist with α close to -1 is in complete disagreement with the rest of the panel, while a panelist with α value equal to +1 is in perfect agreement with the rest of the group. Panelists are then ranked according to the alpha values, and the panelist with smallest α_p is the first to be removed from the analysis. A panelist removal is followed by an attribute selection procedure, detailed in Steps 2 to 6.

Step 2: Apply a multivariate technique on the dataset consisting of sensory attributes

Characterize the relationship among attributes in matrix $\mathbf{X}$ using a multivariate technique, and considering attributes as variables in the analysis. We recommend using PCA on matrix $\mathbf{X}$·PCA outputs of interest are the component weights w_{jr} and the percentage of variance explained by each retained component r ($r = 1, \ldots, R$). The number of retained components R is defined based on the amount of variance explained by them, as in Montgomery, Peck, and Vining (2001).

Step 3: Generate attribute importance indices (z)

Generate an attribute importance index to guide the removal of attributes not relevant for classification purposes, as proposed in Anzanello et al. (2009). Attribute j's index is denoted by z_j $j = 1, \ldots, J$. The higher the value of z_j, the more important the corresponding attribute is for classifying observations into product classes.

The index z_j is generated based on PCA weights, w_{jr}, as in Eq. (2). Attributes with large weights are preferred. According to Duda et al. (2001), such attributes should lead to a better discrimination of observations into classes of products, although exceptions may occur.

$$z_j = \sum_{r=1}^{R} |w_{jr}|, j = 1, \cdots, J \quad (2)$$

Step 4: Classify the dataset using KNN and eliminate irrelevant and noisy attributes

Categorize the $(P \times I \times D)$ dataset observations into I classes of products on all J attributes using KNN, and compute the classification accuracy. Accuracy is defined as the ratio between the number of correct classifications and the number of performed classifications. Parameter k for the KNN algorithm is selected by cross-validation on the sensorial dataset, as in Chaovalitwongse et al. (2007).

Attribute elimination starts by identifying the attribute with the smallest z_j. Remove the selected attribute, perform a new classification using KNN on the $J - 1$ remaining attributes and compute the classification accuracy. This procedure is repeated removing the next attribute with the smallest z_j and applying KNN on the remaining attributes, until there is only one attribute left.

Step 5: Construct an accuracy graph

Construct a graph relating classification accuracy to the number of retained attributes. In case classification accuracy is the only optimization criterion considered, the maximum accuracy indicates the best subset of attributes to be retained for classification. In case of having alternative subsets with identical accuracy values, choose the one with the smallest number of retained attributes.

Step 6: Remove the less consistent panelist and perform a new attribute selection

Remove the panelist with the lowest α_p and repeat Steps 2 to 6 for the sensorial data consisting of the J original attributes and remaining panelists. Add the new accuracy profile to the accuracy graph in Step 5. Note that since the elimination of each panelist leads to a new attribute selection based on the evaluations of remaining panelists, one of the procedure outputs will be a set of accuracy profiles. Repeat this iterative procedure until a user-defined lower bound of remaining panelists is reached.

A final solution is obtained by identifying the maximum global accuracy (peak) on the set of accuracy profiles, as exemplified in Fig. 1. The peak identifies both the best group of panelists and the best subset of attributes to be considered in classification procedures. Pareto Optimal (PO) analysis may also be used to identify

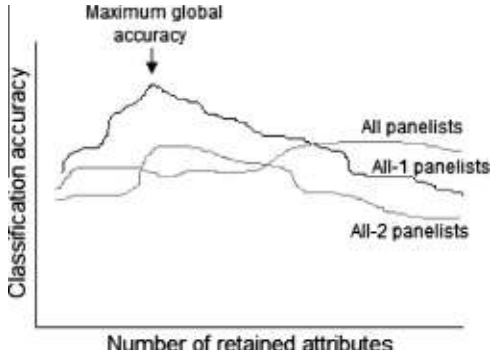


Fig. 1. Accuracy profiles as attributes and panelists are eliminated.

a limited number of solutions that maximize the classification accuracy and minimize the number of retained attributes. Such analysis may be particularly helpful when the graph presents multiple (local) peaks.

Table 1
Sensory attributes evaluated in the experiment.

Appearance attributes	Flavor attributes	Texture attributes
(1) Ratio of gravy to meat	(4) Cooked lean beef	(15) Viscosity of the gravy
(2) Visual thickness of sauce	(5) Beef broth	(16) Springiness of the meat
(*) Hue/value/chroma of the gravy	(6) Organy	(17) Initial cohesiveness of the meat
(*) Hue/value/chroma of the beef	(7) Hydrolyzed vegetable protein	(18) Denseness of the meat
(3) Uniformity of size and shape of the beef	(8) Coagulated beef blood	(19) Firmness of the meat
	(9) Cardboardy	(20) Chewiness of the meat
	(10) Browned	(21) Fibrousness of the meat
	(11) Beef fat	(22) Flaking of the meat
	(12) Salty	(23) Dryness of the bolus
	(13) Metallic feeling flavor	(24) Oily film
	(14) Heat	

Table 2
Alpha values for panelists.

Panelist ID	Alpha
P1	0.8386
P2	0.8205
P3	0.8157
P4	0.8127
P5	0.8104
P6	0.8077
P7	0.7676
P8	0.6959
P9	0.0250

4. Case example

We now apply the suggested method to select discriminant attributes and consistent panelists in a sensory dataset. Twenty-six sensory attributes listed in Table 1 are evaluated by nine trained subjects in a sensory panel; the resulting dataset is available upon request. Attributes are appearance, flavor, and texture-related. Products analyzed are different formulations of stewed beef cubes in gravy, conditioned inside thermostable pouches and processed in a retort. Evaluations are performed following the Spectrum Method, a QDA technique. In the Spectrum Method, products are evaluated individually by each panelist regarding a set of sensory attributes [see Meilgaard, Civille, and Carr (1999) for a description of the method]. Products were prepared through the Combat Ration Advanced Manufacturing Technology Development Program at the Food Manufacturing Technology Facility in Piscataway, NJ, in 1994. Formulations were based on military specifications for beef stew in meal-ready-to-eat pouches. Eight formulations were tested.

In the analyses to follow attributes marked with (*) were not considered due to missing observations from some panelists. Each panelist repeated product evaluations four times on twenty-four attributes; a total of 768 observations were obtained from each panelist. The dimension of matrix **X** containing evaluations from all panelists is (288 × 24). In the analysis reports panelist p is referred to as Pp.

A consistency analysis on panelists using the alpha values in Eq. (1) as consistency index is performed; Table 2 depicts the resulting alpha readings. Panelist 9 (P9) is the least consistent and the first to be removed for attribute selection, followed by P7 and similarly thereafter. In general, panelists display high alpha values, indicating good group consensus in the evaluations performed.

We then apply PCA to the dataset. We set $R = 2$ as the number of components to retain on each PCA performed after a panelist is removed. This number of components explained 63% or more of the variance on the attributes; additional components explained residual portions of the variance, thus not justifying their inclusion.

As for the KKN, parameter $k = 3$ was defined using a five-folded cross-validation. We first define a suitable interval of odd k values [1–11] and use those k's to classify 80% the dataset; the k yielding the maximum classification accuracy is then used to classify the remaining 20% of the observations. That procedure is repeated sev-

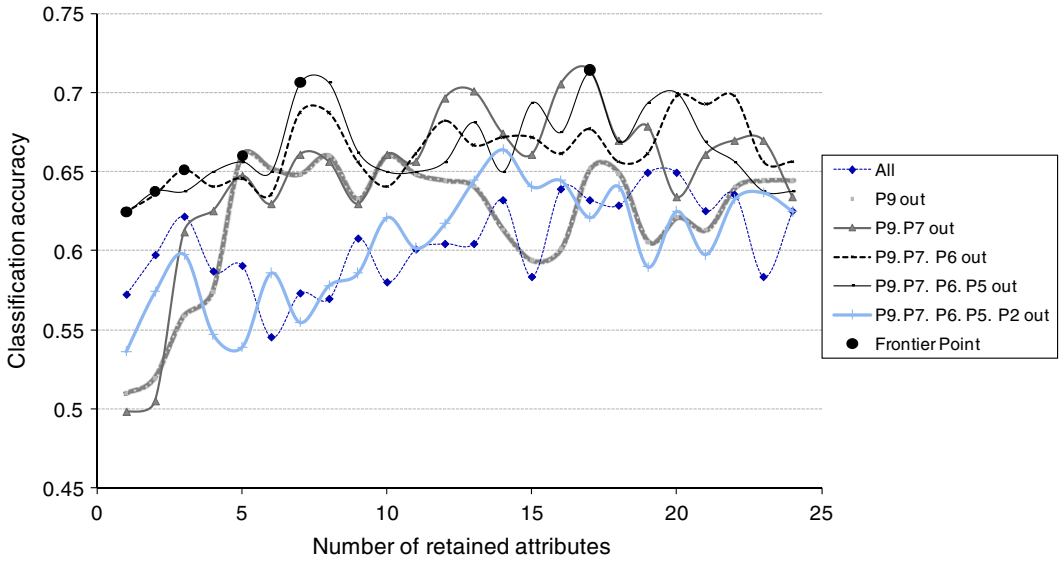


Fig. 2. Accuracy profiles as attributes and panelists are eliminated.

Table 3
Information on pareto optimal frontier points.

Frontier point (FP)	Retained panelist	Retained attribute ID	Classification accuracy
1	P1, P2, P3, P4, P8	16	0.6245
2	P1, P2, P3, P4, P5, P8	16, 18	0.6375
3	P1, P2, P3, P4, P5, P6,P8	16, 19, 18	0.6510
4	P1, P2, P3, P4, P5, P6, P7, P8	22, 11, 21, 16, 18	0.6602
5	P1, P2, P3, P4, P8	20, 18, 16, 19, 22, 17, 4	0.7063
6	P1, P2, P3, P4, P8	20, 18, 16, 19, 22, 17, 4, 21, 3, 6, 23, 10, 12, 11, 1, 2, 8	0.7143

eral times, and the *k* leading to the highest average accuracy on the 20% portion is chosen.

Attribute selection is performed after each panelist elimination, leading to the set of accuracy profiles given in Fig. 2. The elimination of panelists is concluded when a user-defined lower bound of 5 remaining panelists is reached. Frontier points from the PO analysis are identified in the accuracy graph, where FP1 is the extreme left Frontier point. These points are also detailed in Table 3.

Maximum accuracy of 71.4% is obtained when P9, P7, P6 and P5's evaluations are removed from the original dataset, and 17 of the original 24 attributes are retained. That leads to classifications 12% more accurate than using KNN on all 24 original attributes. Retained attributes are 20, 18, 16, 19, 22, 17, 4, 21, 3, 6, 23, 10, 12, 11, 1, 2, and 8 in decreasing order of relevance. Note that classification accuracy using 24 attributes and 9 panelists is lower than the one obtained when 17 attributes and 5 panelists are considered. That is explained by the removal of noisy attributes and inconsistent panelists from the dataset.

Alternative solutions correspond to the frontier points in Table 3; one such solution clearly appears as promising. FP5 displays a situation in which 5 panelists evaluating only 7 attributes attain a classification accuracy of 70.6%. Reducing 70% of the original group of attributes in the sensory experiment without severely compromising classification accuracy is highly desirable if (*i*) stress on panelists is to be minimized, and (*ii*) a low-cost confirmatory sensory panel is to be performed.

5. Conclusion

Reducing the number of attributes to be analyzed in sensory profiling experiments has been object of research in recent years. The objective is to identify a subset of meaningful and non-redundant attributes that allow discrimination of samples tested in the panel. Such optimization is desired to save time and fatigue to panelists, leading to less costly data collection procedures. Many of the attribute selection approaches in the literature are focused on the selection of attributes carrying large amounts of variation. Our scope though is twofold: the identification of attributes that (*i*) lead to the most accurate classification of products and (*ii*) are informative in terms of the proportion of the variance explained in the multivariate dataset.

We propose a method for selecting the important attributes to be used for classification of sensory panel assessed products. The method is: (1) Measure panelists' consistency using a suitable index; (2) Apply PCA on the dataset consisting of sensory attributes; (3) Compute a vector of attribute importance indices based on PCA weights; (4) Classify the sensory dataset using the *k*-Nearest Neighbor (KNN) technique and compute the classification accuracy. Iterate by eliminating the attribute with the lowest importance index, classifying the dataset again, and re-computing the accuracy; (5) Construct an accuracy graph; and (6) Remove the less consistent panelist and perform a new attribute selection, repeating steps 2 to 6.

We apply the proposed method to a descriptive analysis dataset comprised of evaluations performed by 9 panelists on 24 attributes about 8 different product formulations. Maximum accuracy of 71.4% is obtained when four panelists' evaluations are removed from the original dataset, and 17 of the original 24 attributes are retained. An alternative, more parsimonious solution is identified through Pareto Optimal analysis where 5 panelists evaluating only 7 attributes attain a classification accuracy of 70.6%.

Future research includes the development of alternative approaches to select the best subset of attributes for classification. We consider generating and testing other consistency indices to rank attributes and panelists, as well as applying alternative data mining tools, such as Support Vector Machine and Probabilistic Neural Networks. We will also explore methods aimed at selecting attributes for prediction purposes. Generating concise and reliable regression models will enable better product characterization as attributes levels are changed.

Acknowledgements

Dr. Fogliatto's research is supported by CNPq (Grant No. 301380/2008-2). We thank the two reviewers for their valuable comments on an earlier manuscript version.

Appendix

MATLAB CODES FOR ATTRIBUTE SELECTION

```
function clas_PCA_KNN(ts,tr,ta,correct,Knn)
% tr=ts: evaluation data
% ta=correct: formulation classes
% Knn: number of nearest neighbors
[mtr,ntr]=size(tr);
tr1=zscore(tr);
ts1=zscore(ts);

[coefs,scores,variances,t2] = princomp(tr1);
coefs;
WWW1=[sum(abs(coefs(:,1:2))')',(1:ntr)'];
WWW=flipud(sortrows(WWW1,1))

tr1=tr1(:,WWW(:,2));
ts1=ts1(:,WWW(:,2));
WWW2=[WWW(:,2)];
for mm=0:ntr-1
    tr2=tr1(:,1:ntr-mm);
    ts2=ts1(:,1:ntr-mm);
    WWW22=WWW2(1:ntr-mm);

train_patterns1=tr2';
test_patterns1=ts2';
train_targets=ta;

L           = length(train_targets);
Uc          = unique(train_targets);

if (L < Knn),
   error('More neighbors than there are points.')
end
N               = size(test_patterns1, 2);
test_targets    = zeros(1,N);
for i = 1:N,
    dist                                        =    sum(((train_patterns1    -
test_patterns1(:,i)*ones(1,L))).^2);

    [m, indices]    = sort(dist);

    n               = hist(train_targets(indices(1:Knn)), Uc);

    [m, best]       = max(n);

    test_targets(i) = Uc(best);
end
G=test_targets;
H=G'==correct;
ACC_mm=sum(H)/mtr;
ACC(mm+1,1)=ACC_mm;
end
ACC
```

MATLAB CODES FOR PANELISTS' RANKING

```
function panelist(x,npain)
 %npain = number of painelists
 %X = evaluation data
[mx,nx]=size(x);
intpain=mx/npain
for i=1:npain
    if i==1;
    x1=x(1:intpain,:);
   elseif i==npain;
       x1=x(((npain-1)*intpain)+1:npain*intpain,:);
   else i~=1 & i~=npain;
       x1=x(((i-1)*intpain)+1:i*intpain,:);
   end
SC=mean(x1);
[ysize,xsize] = size(SC);
 times=intpain;
for y = 1:ysize,
for rep = 1:times,
out((y - 1) * times + rep,:) = SC(y,:);
end
end
out1=out;
Xc=x1-out1;
t=trace(Xc'*Xc);
theta=1/sqrt(t);
Yc_i=theta*Xc;
YY(intpain*i+1:intpain*i+intpain,:)=Yc_i;
end
   YY=YY(intpain+1:mx+intpain,:)
    s = xlswrite('tempdata.xls',YY)
function tpanelist(YY,npain)
 [mYY,nYY]=size(YY);
intpain=mYY/npain;
for k=1:npain
    if k==1;
    YY1=YY(1:intpain,:);
   elseif k==npain;
       YY1=YY(((npain-1)*intpain)+1:npain*intpain,:);
   else k~=1 & k~=npain;
       YY1=YY(((k-1)*intpain)+1:k*intpain,:);
    end
      for l=1:npain
   if l==1;
    YY2=YY(1:intpain,:);
   elseif l==npain;
       YY2=YY(((npain-1)*intpain)+1:npain*intpain,:);
   else l~=1 & l~=npain;
       YY2=YY(((l-1)*intpain)+1:l*intpain,:);
   end
   s_l=(1+trace(YY1'*YY2))/2
   ss(:,l+1)=s_l
      end
      sss(k+1,:)=ss
end
S=sss(2:k+1,2:l+1)
[V,D] = eig(S);
FEIG=V(:,npain);
FEIGnorm=FEIG/sum(FEIG)
```

```
for k=1:npain
    if k==1;
    YY1=YY(1:intpain,:);
   elseif k==npain;
       YY1=YY(((npain-1)*intpain)+1:npain*intpain,:);
   else k~=1 & k~=npain;
       YY1=YY(((k-1)*intpain)+1:k*intpain,:);
    end
    C_k=FEIGnorm(k)*YY1;
    if k==1;
    CC=C_k;
    else
    CC=CC+C_k;

    end
end
CC
for k=1:npain
    if k==1;
    YY1=YY(1:intpain,:);
   elseif k==npain;
       YY1=YY(((npain-1)*intpain)+1:npain*intpain,:);
   else k~=1 & k~=npain;
       YY1=YY(((k-1)*intpain)+1:k*intpain,:);
    end
    alpha_k=trace(YY1'*CC)/sqrt(trace(CC'*CC))
```

References

Anzanello, M. J., Albin, S. L., & Chaovalitwongse, W. (2009). Selecting the best variables for classifying production batches into two quality classes. *Chemometrics and Intelligent Laboratory Systems, 97*(2), 111–117.

Azapagic, A. (1999). Life cycle assessment and its application to process selection, design and optimization. *Chemical Engineering Journal, 73*(1), 1–21.

Camara, J., Alves, M., & Marques, J. (2006). Multivariate analysis for the classification and differentiation of Madeira wines according to the main grape varieties. *Talanta, 68*, 1512–1521.

Capron, X., Smeyers-Verbeke, J., & Massart, D. (2007). Multivariate determination of the geographical origin of wines from four different countries. *Food Chemistry, 101*, 1585–1597.

Carbonell, L., Izquierdo, L., & Carbonell, I. (2007). Sensory analysis of Spanish mandarin juices: Selection of attributes and panel performance. *Food Quality and Preference, 18*, 329–341.

Chabanet, C. (2000). Statistical analysis of sensory profiling data. Graphs for presenting results (PCA and ANOVA). *Food Quality and Preference, 11*(1–2), 159–162.

Chaovalitwongse, W., Fan, Y., & Sachdeo, C. (2007). On the time series k-nearest neighbor classification of abnormal brain activity. *IEEE Transactions on System and Man Cybernetics A, 37*(6), 1005–1016.

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multi objective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation, 6*(2), 182–197.

Deb, K., Thiele, L., Laumanns, M., & Zitzler, E. (2002). Scalable multi-objective optimization test problems. *Proceedings of the 2002 Congress on Evolutionary Computation, 1*, 825–830.

Derde, M., & Massart, D. (1986). Supervised pattern recognition: The ideal method? *Analytica Chimica Acta, 184*, 33–51.

Dijksterhuis, G., Frost, M. B., & Byrne, D. V. (2002). Selection of a subset of variables: Minimization of Procrustes loss between a subset and the full set. *Food Quality and Preference, 13*, 89–97.

Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification* (2nd ed.). New York: Wiley-Interscience.

Fogliatto, F. S., Albin, S. L., & Tepper, B. J. (1999). A hierarchical approach to optimizing descriptive analysis multiresponse experiments. *Journal of Sensory Studies, 14*(4), 443–465.

Gauchi, J., & Chagnon, P. (2001). Comparison of selection methods of exploratory variables in PLS regression with application to manufacturing process data. *Chemometrics and Intelligent Laboratory Systems, 58*, 171–193.

Granitto, P., Biasioli, F., Endrizzi, I., & Gasperi, F. (2008). Discriminant models based on sensory evaluations: Single assessors versus panel average. *Food Quality and Preference, 19*(6), 589–595.

Granitto, P. M., Gasperi, F., Biasioli, F., Trainotti, E., & Furlanello, C. (2007). Modern data mining tools in descriptive sensory analysis: A case study with a Random forest approach. *Food Quality and Preference, 18*, 681–689.

Guo, Q., Wu, W., Massart, D., Boucon, C., & Jong, S. (2002). Feature selection in principal component analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems, 61*, 123–132.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Horn, J., Nafpliotis, N., & Goldberg, D. (1994). A niched pareto genetic algorithm for multiobjective optimization. In *Proceedings of the First IEEE Conference on Evolutionary Computation, IEEE World Congress on Computational Intelligence*, 1, 82–87.

Jansen, J., Hoefsloot, H., Greef, J., Timmerman, M., Westerhuis, J., & Smilde, J. (2005). ASCA: Analysis of multivariate data obtained from an experimental design. *Journal of Chemometrics, 19*(9), 469–481.

Kudo, M., & Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition, 33*, 25–41.

Lavit, C., Escoufier, Y., Sabatier, R., & Traissac, P. (1994). The ACT (STATIS method). *Computational Statistics & Data Analysis, 18*, 97–119.

Ledauphin, S., Hanafi, M., & Qannari, E. M. (2006). Assessment of the agreement among the subjects in fixed vocabulary profiling. *Food Quality and Preference, 17*(3–4), 277–280.

Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering, 17*(4), 491–502.

Luciano, G., & Naes, T. (2009). Interpreting sensory data by combining principal component analysis and analysis of variance. *Food Quality and Preference, 20*(3), 167–175.

Mallet, Y., De Vel, O., & Coomans, D. (1998). Integrated feature extraction using adaptive wavelets. In: H. Liu & H. Motoda, Feature extraction, construction and selection: A Data mining perspective, 175–189.

Marini, F., Bucci, R., Magri, A., & Magri, A. (2006). Authentication of Italian CDO wines by class-modeling techniques. *Chemometrics and Intelligent Laboratory Systems, 84*, 164–171.

Meilgaard, M., Civille, G. V., & Carr, B. T. (1999). *Sensory evaluation techniques* (3rd ed.). Boca Raton: CRC Press.

Montgomery, D., Peck, E., & Vining, G. (2001). *Introduction to linear regression analysis*. New York: John Wiley.

Murray, J. M., Delahunty, C. M., & Baxter, I. A. (2001). Descriptive sensory analysis: Past, present and future. *Food Research International, 34*(6), 461–471.

Rason, J., Marin, J., Dufour, E., & Lebecque, A. (2007). Diversity of the sensory characteristics of traditional dry sausages from the centre of France. Relation with regional manufacturing practice. *Food Quality and Preference, 18*(3), 517–530.

Rebolo, S., Pena, R., Latorre, M., Botana, A., & Herrero, C. (2000). Characterisation of Galician (NW Spain) Ribeira Sacra wines using pattern recognition analysis. *Analytica Chimica Acta, 417*, 211–220.

Ripley, B. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.

Sahmer, K., & Qannari, E. M. (2008). Procedures for the selection of a subset of attributes in sensory profiling. *Food Quality and Preference, 19*, 141–145.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics*. New York: Harper Collins College Publishers.

Taboada, H., & Coit, D. (2007). Data clustering of solutions for multiple objective system reliability optimization problems. *Quality Technology & Quantitative Management Journal, 4*, 35–54.

Taboada, H., & Coit, D. (2008). Multi-objective scheduling problems: Determination of pruned Pareto sets. *IIE Transactions, 40*, 552–564.

Urtubia, A., Perrez-Correa, J., Soto, A., & Pszczolkowski, P. (2007). Using data mining techniques to predict industrial wine problem fermentation. *Food Control, 18*, 1512–1517.

Weiss, S., Apte, C., Dameray, D., Johnson, D., Ples, F., Goetz, T., et al. (1999). Maximizing text-mining performance. *IEEE Intelligent Systems, 14*(4), 63–69.

Westad, F., Hersleth, M., Lea, P., & Martens, H. (2003). Variable selection in PCA in sensory descriptive and consumer data. *Food Quality and Preference, 14*, 463–472.

Wold, S., & Sjostrom, M. (1977). A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. In B. R. Kowalski (Ed.), *Chemometrics, Theory and application, ACS Symposium series. 52* (pp. 243–282). Washington, DC: American Chemical Society.

Wu, X., Kumar, V., Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., et al. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems, 14*(1), 1–37.

Zitzler, E., & Thiele, L. (1999). Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach. *IEEE Transactions on Evolutionary Computation, 3*(4), 257–271.