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

Victor Leonardo Cervo

SELEÇÃO DE VARIÁVEIS PARA CLUSTERIZAÇÃO ATRAVÉS DE ÍNDICES DE IMPORTÂNCIA DAS VARIÁVEIS E ANÁLISE DE COMPONENTES PRINCIPAIS

Porto Alegre
2013

Victor Leonardo Cervo

Seleção de variáveis para clusterização através de índices de importâncias das variáveis e Análise de Componentes Principais

Dissertação 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 Mestre em Engenharia de Produção, modalidade Acadêmica, na área de concentração em Sistemas de Qualidade.

Orientador: Michel Jose Anzanello, *Ph.D.*

Porto Alegre

2013

Victor Leonardo Cervo

Seleção de variáveis para clusterização através de índices de importâncias das variáveis e Análise de Componentes Principais

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

_____

Prof. Michel Jose Anzanello, *Ph.D.*
Orientador PPGEP/UFRGS

_____

Prof. José Luis Duarte Ribeiro, Dr.
Coordenador PPGEP/UFRGS

Banca Examinadora:

Prof. Fernando Hepp Pulgati, Dr. (DEST/UFRGS)

Profa. Liane Werner, Dra. (PPGEP/UFRGS)

Prof. Marcelo Farenzena, Dr. (DEQUI/UFRGS)

AGRADECIMENTOS

Agradeço, primeiramente, a Deus, pela oportunidade maravilhosa de poder trilhar os caminhos que trilhei e pela determinação em seguir em frente.

À minha esposa, Josiane, e filho, João Vitor, por tudo o que representam na minha vida: a felicidade, o amor, a amizade, a lealdade e o aprendizado.

Aos meus pais, Ivo e Maria Ângela, pelo incentivo aos estudos.

A Antonia e Regina, pelo apoio desde o início dessa jornada.

Ao meu orientador, Prof. Michel José Anzanello, *Ph.D.*, pela paciência, dedicação, e amizade.

Aos amigos que fiz nessa empreitada: Diego, Eduardo, Marco e Thiago pelo apoio e convivência.

Aos professores e colegas do Programa de Pós-Graduação em Engenharia de Produção, por todo o apoio e incentivo que recebi nos últimos dois anos.

CERVO, Victor Leonardo ***Seleção de variáveis para clusterização através de índices de importância de variáveis e Análise de Componentes Principais,*** 2013. Dissertação (Mestrado em Engenharia) - Universidade Federal do Rio Grande do Sul, Brasil.

RESUMO

A presente dissertação propõe novas abordagens para seleção de variáveis com vistas à formação de grupos representativos de observações. Para tanto, sugere um novo índice de importância das variáveis apoiado nos parâmetros oriundos da Análise de Componentes Principais (APC), o qual é integrado a uma sistemática do tipo *forward* para seleção de variáveis. A qualidade dos agrupamentos formados é medida através do *Silhouette Index*. Um estudo de simulação é projetado para avaliar a robustez e o desempenho da sistemática proposta em dados com diferentes níveis de correlação, ruído e número de observações a serem clusterizadas. Na sequência, é apresentada uma versão modificada da sistemática original, a qual utiliza funções *kernel* para remapeamento dos dados com vistas ao incremento da qualidade de clusterização e redução das variáveis retidas para formação dos agrupamentos. A versão modificada é aplicada em 3 bancos de dados da indústria química, aumentando a qualidade da clusterização medida pelo SI médio em 150% e utilizando em torno de 6% das variáveis originais.

Palavras-chave: Seleção de variáveis, análise de clusterização, análise de componentes principais, funções de remapeamento.

CERVO, Victor Leonardo ***Clustering variable selection through variable importance indices and Principal Component Analysis,*** 2013. Thesis (Master in Engineering) - Federal University of Rio Grande do Sul, Brazil.

ABSTRACT

This thesis proposes new approaches for variable selection aimed at forming representative groups of observations. For that matter, we suggest a new variable importance index based on parameters derived from the Principal Component Analysis (PCA), which is integrated to a forward procedure for variable selection. The quality of clustering procedure is assessed by the Silhouette Index. A simulation study is designed to evaluate the robustness of the proposed method on different levels of variable correlation, noise and number of observations to be clustered. Next, we modify the original method by remapping observations through *kernel* functions tailored to improving the clustering quality and reducing the retained variables. The modified version is applied to 3 databases related to chemical processes, increasing the quality of clustering measured by SI on average 150%, while using around 6% of the original variables.

Keywords: Variable selection, clustering analysis, principal component analysis, *kernel* functions.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 Introdução

1.1 Considerações Iniciais

A separação de observações em grupos distintos tem sido alvo de estudos em diversas áreas de pesquisa, visto que permite melhor compreender as relações existentes entre as observações em estudo (Hair *et al.*, 1995). As técnicas de clusterização, especificamente, procuram promover essa separação de forma que os grupos formados – *clusters* – sejam distintos (representativos), e que cada *cluster* contenha observações similares entre si e distintas em relação a observações inseridas em outros *clusters* (Kaufmann e Rousseeuw, 2005). Dentre os diversos métodos de clusterização existentes na literatura, um dos mais utilizados é o *k-means*, um método de particionamento em *k clusters*, baseado na definição de elementos centrais de *clusters*, chamados centroides (STEINLEY, 2006).

O objeto de estudo da análise de clusterização é o conjunto de observações, geralmente descrito por um conjunto de variáveis. Devido à evolução tecnológica, mais variáveis podem ser monitoradas simultaneamente, elevando de forma expressiva a quantidade de dados disponíveis para análise. Num primeiro instante, pode-se pensar que uma maior quantidade de dados, representados por um volume elevado de variáveis descritivas de observações, seja preferível para a formação de grupos representativos. Contudo, diversos estudos sugerem que *clusters* mais consistentes são obtidos pela utilização de um subconjunto reduzido das variáveis originais (MILLIGAN, 1980; LI *et al*.,2008; MAUGIS *et al*., 2009; ANZANELLO; FOGLIATTO, 2011).

Com o objetivo de determinar esse subconjunto de variáveis, diversas sistemáticas têm sido propostas. A maioria delas explora a possibilidade de definir diferentes níveis de importância às variáveis através da atribuição de diferentes pesos (Fowlkes *et al*., 1988; Gnanadesikan *et al*., 1995; Friedman e Meulman, 2004; Huang *et al*., 2005; Steinley e Brusco, 2008). Alguns estudos, no entanto, mostram que a atribuição de pesos às variáveis irrelevantes atrapalha o processo de recuperação de *clusters*, uma vez que a influência de tais variáveis ponderadas, mesmo em menor escala, continua a inserir informações ruidosas no processo. Com isso, sugere-se a exclusão desses efeitos indesejáveis através da seleção de variáveis, onde uma variável tida como irrelevante é efetivamente removida do processo de clusterização (GNANADESIKAN *et al*., 1995; BRUSCO; CRADIT, 2001; BRUSCO, 2004; LI *et al*., 2008; MAUGIS *et al*., 2009; ANZANELLO; FOGLIATTO, 2011).

Dentre as muitas sistemáticas propostas para seleção de variáveis, merecem destaque as que utilizam ferramentas multivariadas como MANOVA (*Multivariate Analysis of Variance*), (Gnanadesikan *et al*., 1995), regressão PLS (*Partial Least Squares*) (Anzanello *et al*., 2009) e Análise de Componentes Principais (ACP) (Steinley e Brusco, 2008). Nessa linha, a ACP surge como uma ferramenta enraizada em premissas matemáticas simples, tratando-se da representação dos dados em uma nova estrutura, obtida através de transformações lineares nos dados e fortemente fundamentada em informações sobre variância das variáveis (JOLLIFFE, 2002; ANDERSON, 2003).

A fim de realizar a seleção de variáveis de forma sistemática, o tema da presente dissertação consiste na seleção de variáveis para clusterização através de índices de importância das variáveis, gerados a partir dos parâmetros da ACP. Esta dissertação é composta por três artigos que abordam a seleção de variáveis de clusterização. No primeiro artigo é proposta uma sistemática de seleção de variáveis para clusterização com base em um novo índice de importância das variáveis. O segundo artigo avalia o método proposto no primeiro artigo frente a variações nos níveis de correlação, ruído na coleta dos dados e proporção dos dados disponíveis para análise, através de experimentos de simulação. O terceiro artigo propõe a utilização de funções *kernel* com vistas ao remapeamento das observações, a fim de inserir na análise relações não lineares que possam conduzir a agrupamentos mais precisos.

1.2 Objetivos

O objetivo principal do trabalho é propor metodologias de seleção de variáveis de clusterização a fim de obter agrupamentos representativos de observações.

Como objetivos específicos surgem:

- Apresentar a fundamentação teórica da análise de clusterização sob a ótica de seleção de variáveis;
- Integrar a ACP a técnicas de clusterização com vistas à seleção de variáveis;
- Desenvolver um novo índice de importância de variáveis a partir dos parâmetros oriundos da ACP, o qual serve como guia para identificação das variáveis mais relevantes;
- Avaliar a robustez do método proposto através de um experimento de simulação abordando fatores considerados relevantes para seleção de variáveis;

- Avaliar os benefícios do remapeamento dos dados originais através de funções *kernel* na sistemática proposta em termos de qualidade dos agrupamentos formadas e percentual de variáveis retidas.

1.3 Justificativa do Tema e dos Objetivos

O desenvolvimento de sistemáticas para seleção de variáveis de clusterização com vistas à formação de grupos representativos encontra respaldo prático e teórico.

Diversas aplicações práticas apoiam-se na coleta e análise de um grande número de variáveis e de observações. Tais aplicações incluem a área médica, com o intuito de prever diagnósticos (Wolberg *et al*., 1993; Detrano *et al*., 1989), áreas sociais, a fim de estudar aspectos demográficos (Meek *et al*., 2002), gestão de produção, com o propósito de facilitar a programação de manufatura de famílias de produtos, e ciências aplicadas em geral, com vistas a reconhecer padrões entre os dados observados (van Breukelen *et al*., 1998). Em tais aplicações, a formação de grupos consistentes com base em um subconjunto formado por variáveis relevantes permite que conclusões acerca de algumas observações possam ser estendidas às demais observações de um mesmo *cluster*. Percebe-se então a redução de esforços e custos de coleta de dados, e maior agilidade nas tomadas de decisões acerca das famílias (e não das observações individuais), entre outros benefícios.

Em termos teóricos, percebe-se o aumento de pesquisas devotadas à identificação das variáveis com maior capacidade de formação de agrupamentos representativos (Gauchi e Chagnon, 2001; Anzanello *et al*., 2009; Maugis *et al*., 2009; Anzanello e Fogliatto, 2011). Tais abordagens têm se apoiado em ferramentas multivariadas de diversas complexidades, permitindo a geração de métodos que normalmente apresentam desempenho satisfatório em aplicações específicas, mas não são generalizáveis a outras áreas (Fowlkes *et al*., 1988; Gnanadesikan *et al*., 1995; Brusco e Cradit, 2001; Brusco, 2004, Li *et al*., 2008). Tendo-se em vista que não existe uma sistemática unânime para identificação das variáveis mais relevantes, justifica-se o desenvolvimento de abordagens mais eficientes e preferencialmente simples para a seleção de variáveis para clusterização, tema dos três artigos dessa dissertação.

1.4 Procedimentos Metodológicos

O método de pesquisa utilizado nessa dissertação pode ser classificado como de natureza aplicada, tendo em vista que objetiva gerar conhecimentos voltados à solução de problemas existentes. Tem objetivo exploratório, pois a partir da análise das hipóteses construídas, busca a resolução de um problema de cunho prático. Apresenta-se ainda como uma abordagem quantitativa, em virtude de utilizar ferramentas matemáticas para análise de um estudo de caso (GIL, 2002; SILVA; MENEZES, 2005).

1.5 Estrutura da Dissertação

A dissertação está organizada em 5 capítulos. O primeiro capítulo traz a introdução do trabalho, apresentando e justificando o tema e os objetivos, e descreve os procedimentos metodológicos adotados. As delimitações da pesquisa vêm a seguir, de forma que a estrutura do trabalho encerra o capítulo.

O segundo capítulo traz o primeiro artigo, o qual apresenta uma revisão da literatura sobre o ferramental utilizado: análise de clusterização e análise de componentes principais. Além disso, apresentam-se estudos relacionados ao tema da seleção de variáveis para clusterização, alguns dos quais integrando as duas ferramentas estudadas. Apresenta-se uma sistemática de seleção de variáveis para clusterização com base em um novo índice de importância das variáveis, gerado a partir de parâmetros da ACP. O método proposto é aplicado em 3 bancos de dados reais, e a qualidade da clusterização é avaliada através do *Silhouette Index* (SI).

O terceiro capítulo apresenta o segundo artigo, que visa avaliar o desempenho do método proposto no primeiro artigo em dados com diferentes níveis de correlação, ruído e proporção de observações disponíveis para análise. São brevemente analisados trabalhos envolvendo simulação no contexto de seleção de variáveis. Um projeto de simulação é estruturado de forma que os níveis do experimento sejam avaliados, a fim de observar sua influência sobre a qualidade da clusterização e o percentual de variáveis retidas.

O quarto capítulo traz o terceiro artigo, o qual introduz uma ferramenta de mapeamento dos dados de entrada para um novo espaço de representação, baseado na teoria das funções *kernel*. É apresentada uma revisão de conteúdos pertinentes sobre essas funções e mapeamento de dados. São apresentados trabalhos que envolvem *kernels* na análise de clusterização e ressaltadas as diferentes abordagens existentes. Uma metodologia de mapeamento dos dados é então implementada com a finalidade de inserir não-linearidades na

análise. A metodologia é comparada à proposta no primeiro artigo ao ser aplicada em 3 bancos de dados da indústria química.

O quinto e último capítulo apresenta a conclusão do trabalho, com as avaliações dos principais resultados obtidos frente aos objetivos inicialmente traçados e as sugestões para estudos futuros.

1.6 Delimitações do Estudo

O presente estudo apresenta as seguintes restrições:

- Dentre todos os algoritmos disponíveis na literatura para fins de clusterização, será estudado somente o *k-means*.
- Não é objetivo desse trabalho propor um novo algoritmo de clusterização, mas sim uma nova sistemática de seleção de variáveis a ser utilizada em conjunto com um algoritmo já existente (*k-means*).
- O foco do trabalho restringe-se à clusterização de observações (método não supervisionado), não abordando técnicas de classificação (métodos supervisionados).

1.7 Referências Bibliográficas

Anderson, T.W., 2003. **An Introduction to Multivariate Statistical Analysis** third ed. John Wiley & Sons, Inc. Hoboken, New Jersey.

Anzanello, M.J., Albin, S.L., Chaovalitwongse, W., 2009. Selecting the Best variables for classifying production batches into two quality levels. **Chemometrics and Intelligent Laboratories Systems** 97 (2), 111-117.

Anzanello, M.J., Fogliatto, F.S., 2011. Selecting the best clustering variables for grouping mass-customized products involving workers' learning. **International Journal of Production Economics** 130 (2), 268-276.

Brusco, M.J., 2004. Clustering binary data in the presence of masking variables. **Psychological Methods**, 9, 510-523.

Brusco, M.J., Cradit, J.D., 2001. A variable-selection heuristic for k-means clustering. **Psychometrika** 66 (2), 249-270.

Detrano, R., Jarosi, A.,Steinbrunn, W., Pfisterer, M.m Schmid, J., Sandhu, S., Guppy, K., Lee, S., Froelicher, V., 1989. International application of a new probability algorithm for the diagnosis of coronary artery disease. **American Journal of Cardiology** 64, 304-310.

Fowlkes, E., Gnanadesikan, R., Kettenring, J., 1988. Variable selection in clustering. **Journal of Classification** 5 (2), 205-228.

Friedman, J.H., Meulman, J.J., 2004. Clustering objects on subsets of attributes (with discussion). **Journal of the Royal Statistical Society**, Series B 66, 815-849.

Gauchi, J.P., Chagnon, P., 2001. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. **Chemometrics Intelligent Laboratory Systems** 58, 171-193.

Gil, A.C., 2002. **Como elaborar projetos de pesquisa**. 4 ed. São Paulo – Atlas.

Gnanadesikan, R., Kettenring, J., Tsao, S., 1995. Weighting and selection of variables for cluster analysis. **Journal of Classification** 12 (1), 113-136.

Hair, J., Anderson, R., Tatham, R., Black, W., 1995. **Multivariate Data Analysis with Readings** fourth ed. Prentice-Hall Inc., New Jersey.

Huang, J.Z., Ng, M.K., Rong, H. Li, Z., 2005. Automated variable weighting in k-means type clustering. **IEEE Transactions on Pattern Analysis and Machine Intelligence** 27 (5),657-668.

Jolliffe, I.T., 2002. **Principal Component Analysis** second ed. Springer-Verlag New York.

Kaufman, L., Rousseeuw, P., 2005. **Finding Groups in Data: an Introduction to Cluster Analysis**. Wiley Interscience, New Jersey.

Li, Y., Dong, M., Hua, J., 2008. Localized feature selection for clustering. **Pattern Recognition Letters** 29 (1), 10-18.

Maugis, C., Celeux, G., Martin-Magniette, M., 2009. Variable selection for clustering with Gaussian mixture models. **Biometrics** 65 (3), 701-709.

Meek, C., Thiesson, B., Heckerman, D., 2002. The learning-curve sampling method applied to model-based clustering. **Journal of Machine Learning Research** 2, 397-418.

Milligan, G., 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. **Psychometrika** 45, 325-342.

Silva, E.L., Menezes, E.M., 2005. **Metodologia de pesquisa e elaboração de dissertação**. Florianópolis: Laboratório de ensino da Universidade Federal de Santa Catarina.

Steinley, D., 2006. *K*- means clustering: A half centuty synthesis. **British Journal of Mathematical and Statistical Psychology** 59, 1-34.

Steinley, D., Brusco, M.J., 2008. A new variable weighting and selection procedure for K-means cluster analysis. **Multivariate Behavioral Research** 43 (1), 77-108.

van Breukelen, M., Duin, R.P.W., Tax, D.M.J., den Hartog, J.E., 1998. Handwritten digit recognition by combined classifiers. **Kybernetica** 34 (4), 381-386.

Wolberg, W.H., Street, W.N., Mangasarian, O.L., 1994. Machine learning techniques do diagnose breast cancer from fine-niddle aspirates. **Cancer Letters** 77, 163-171.

2 Primeiro Artigo: Sistemática de seleção de variáveis para clusterização baseada em análise de componentes principais

Victor Leonardo Cervo

Michel Jose Anzanello

Artigo enviado para publicação na revista Gestão e Produção

Resumo

A análise de clusterização é uma ferramenta fundamental em diversas áreas de conhecimento, uma vez que a correta identificação da estrutura inerente às observações de uma amostra dificilmente é uma tarefa trivial. Nesse contexto, busca-se a separação dessas observações em *clusters* distintos entre si, e com alto grau de similaridade dentro de cada *cluster*. O objetivo deste trabalho é agrupar observações utilizando uma metodologia de seleção de variáveis, a qual se baseia em análise de componentes principais. Quando aplicado em três bancos de dados distintos, o método aprimorou a qualidade do agrupamento de observações, avaliado através do *Silhouette Index*, em aproximadamente 100% em média, utilizando em torno de 9% das variáveis originalmente disponíveis.

Palavras-chave: análise de clusterização, seleção de variáveis, análise de componentes principais

Framework for variable selection in clustering based on principal components analysis

Abstract

The clustering analysis is an essential tool in many fields of knowledge, since the correct identification of the structure inherent to observations of a sample is hardly a trivial task. In this context, we seek the separation of these observations into distinct *clusters* among themselves, and with a high degree of similarity within each cluster. The objective of this work is to group observations using a variable selection methodology, which is based on principal component analysis. Applied in three different databases, the method could improve

the grouping of individuals as measured by the Silhouette Index, by 100% on average using around 9% of the variables originally available.

Key-words: clustering analysis, variable selection, principal components analysis

2.1 Introdução

Técnicas de clusterização têm sido utilizadas com os mais variados propósitos em diferentes áreas do conhecimento. A classificação de objetos em categorias vem sendo realizada com o objetivo de melhor compreender os fenômenos naturais que relacionam esses objetos. Brusco *et al*. (2012) apontam técnicas emergentes de clusterização, suas indicações e limitações. Os agrupamentos gerados possibilitam melhor alocação de recursos em contextos produtivos, além de facilitarem procedimentos de coleta de dados (ou seja, a coleta de informações de um membro pertencente a determinado grupo pode representar o comportamento de todos os integrantes daquele grupo).

Intuitivamente, pode-se considerar que um maior volume de dados, representado por elevado número de variáveis descritivas da observação a ser alocada a um grupo, conduza a melhores agrupamentos. No entanto, percebe-se que a inclusão de variáveis irrelevantes e ruidosas no procedimento de clusterização reduz a qualidade dos agrupamentos (Anzanello e Fogliatto, 2011). Maugis *et al*. (2009) afirmam que a estruturação de agrupamentos tipicamente depende de um grupo reduzido de variáveis, e que algumas podem atrapalhar nessa identificação. Em sistemas produtivos, tal imprecisão pode resultar na alocação inadequada de recursos a modelos de produtos que não demandam etapas de processamento similares. Em sistemas médicos, agrupamentos inadequados podem mascarar possíveis portadores de certas doenças como pacientes que não têm tendência a desenvolver essas enfermidades.

A escolha das melhores variáveis para clusterização pode ser operacionalizada, de forma genérica, através da atribuição de pesos às variáveis. Esses pesos determinam com que intensidade as variáveis afetam a estruturação dos agrupamentos. Vários autores vêm realizando estudos nesse campo, dentre os quais destacam-se Gnanadesikan *et al*. (1995), Honda *et al*. (2009), Steinley e Brusco (2008b), Brusco e Cradit (2001) e Fowlkes *et al*. (1988). O problema da seleção de variáveis, por sua vez, pode ser entendido como uma classe especial de problemas de atribuição de pesos, na qual se atribui peso zero às variáveis

irrelevantes e ruidosas, enquanto que às variáveis relevantes é atribuído peso unitário (Brusco e Cradit, 2001). Abordagens alternativas para seleção de variáveis em problemas de clusterização e de classificação têm se apoiado ainda em ferramentas multivariadas, como regressão PLS (Anzanello *et al*., 2009) e ANOVA multivariada (Gnanadesikan *et al*., 1995). Com propósitos semelhantes, Steinley e Brusco (2008b) utilizam informação relativa à variância das variáveis para selecionar as mesmas, a qual constitui-se em uma abordagem mais alinhada à proposta deste estudo.

Este artigo propõe um método para seleção das variáveis mais relevantes com propósitos de clusterização utilizando análise de componentes principais (ACP). Para tanto, a ACP é aplicada sobre os dados (variáveis descrevendo observações a serem agrupadas), e um índice de importância das variáveis de clusterização é gerado. Este índice, baseado nos parâmetros oriundos da ACP, é utilizado para hierarquizar a variância das variáveis para clusterização; variáveis com maiores índices de importância são tidas como mais relevantes em procedimentos de classificação e clusterização (Duda *et al*., 2001). Um procedimento iterativo de clusterização é então iniciado valendo-se da variável mais relevante, e a qualidade dos agrupamentos formados é medida através do *Silhouette Index* (SI), fazendo-se a média do valor obtido por todas as observações. Esta métrica representa o quanto a alocação de uma observação a um grupo é mais adequada do que a alocação no grupo vizinho mais próximo. Na sequência, a segunda variável com maior índice de importância é inserida no subconjunto de variáveis para clusterização, e a clusterização é rodada com base nesse subconjunto. Esse processo é repetido até que todas as variáveis sejam testadas como variáveis de clusterização através de um procedimento *forward* de inclusão de variáveis, sendo que o SI médio das observações é recalculado a cada nova clusterização (gerando um perfil de qualidade de clusterização com a inserção das variáveis no procedimento). A sistemática é então operacionalizada para diferentes números de *clusters*, permitindo identificar o melhor número de agrupamentos a ser formado. O conjunto de variáveis responsável pelo maior SI médio é recomendado para futuras clusterizações.

A principal contribuição do método proposto está na proposição de uma sistemática apoiada em técnicas multivariadas para seleção de variáveis com vistas à formação de agrupamentos. A proposição de um novo índice de importância, o qual se baseia nos parâmetros gerados pela ACP, também apresenta vantagens frente a outros métodos de seleção propostos por conta de sua simplicidade e praticidade.

O artigo está organizado conforme segue. A Seção 2.2 apresenta uma breve revisão da literatura sobre clusterização e ACP. A Seção 2.3 descreve o método proposto. Os resultados, assim como a discussão acerca dos mesmos, são apresentados na Seção 2.4. A Seção 2.5 encerra o artigo, trazendo as conclusões e os direcionamentos futuros.

2.2 Fundamentação teórica

Esta seção apresenta os fundamentos das ferramentas em que a sistemática proposta se apoia: técnicas de clusterização e análise de componentes principais.

2.2.1 Análise de clusterização

A análise de clusterização busca primordialmente realizar a alocação de observações, as quais são descritas por características (variáveis), em grupos, de forma que as similaridades sejam grandes entre observações dentro de um mesmo *cluster* (Hair *et al*., 1995). Cada grupo de observações deve, assim, apresentar grande semelhança interna, ao mesmo tempo em que, se a separação dessas for adequada, as observações de um *cluster* devem ser bastante diferentes das inseridas em outro (AGARD; PENZ, 2009).

Existem dois conjuntos de algoritmos para realizar a clusterização de um conjunto de observações: hierárquicos e não hierárquicos (Hair *et al*., 1995). Os algoritmos hierárquicos baseiam-se na construção de uma hierarquia entre os indivíduos, sendo esta graficamente representada através de uma estrutura chamada dendograma, semelhante a uma árvore. Os *clusters* formados são cortes realizados nos ramos dessa árvore, sendo essa técnica usualmente utilizada para estimar um número adequado de agrupamentos a serem gerados. Esse tipo de clusterização não será abordado neste estudo. Os algoritmos não hierárquicos, dentre os quais se destaca o *k-means*, não envolvem a construção de estruturas do tipo árvore; tais técnicas agrupam as observações em *k clusters*, sendo este um valor previamente conhecido para o algoritmo, a partir da definição de centroides, que são os elementos centrais de cada cluster (Hair *et al*., 1995). Esses centroides são usualmente escolhidos de forma aleatória pelos algoritmos de clusterização.

Matematicamente, as observações são alocadas a um determinado *cluster* de forma a minimizar a soma global das distâncias entre as observações dentro de um *cluster* e o centroide desse *cluster*. Existem diversas métricas para calcular essa distância, sendo a distância euclidiana a mais comum. Considere um sistema de coordenadas cartesiano, com eixos x e y; a distância euclidiana simples entre dois pontos, $P_1(X_1,Y_1)$ e $P_2(X_2,Y_2)$, é calculada conforme a Eq. (2.1):

$$distancia = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \quad (2.1)$$

Além da distância Euclidiana, existem outras formas de medir a similaridade de observações a serem inseridas em grupos: a medida *city-block*, por exemplo, consiste na soma das diferenças absolutas; outra forma de medir a similaridade pode utilizar a correlação entre as variáveis. A correlação, ao contrário das medidas baseadas em distâncias, não considera a magnitude dos valores, mas sim os padrões desses; tal abordagem é pouco utilizada devido ao enfoque da maioria das aplicações de análise de clusterização, que concentra-se na magnitude dos objetos (HAIR *et al*., 1995).

A avaliação do desempenho da clusterização pode ser realizada através do *Silhouette Index* (SI), o qual avalia o quanto uma observação é semelhante às outras observações inseridas em seu *cluster*, comparado com observações inseridas em outros *clusters* (Kaufman e Rousseeuw, 2005). Cada observação apresenta um SI_n, o qual varia no intervalo [-1;1]; n representa a observação que está sendo avaliada, onde n:1,...,N. Valores de SI_n próximos a 1 indicam que a distância, ou dissimilaridade, entre a observação e outras observações alocadas em outros *clusters* é pequena; assim, considera-se que a observação foi corretamente alocada ao *cluster* atual. Valores próximos a -1 indicam que a observação foi provavelmente alocada a um *cluster* inadequado. Valores intermediários, ou seja, próximos a 0, indicam observações que não pertencem a um *cluster* ou outro. O SI_n é calculado de acordo com a Eq. (2.2) (ANZANELLO; FOGLIATTO, 2011):

$$\mathrm{SI}_n = \frac{b(n) - a(n)}{\max\{b(n), a(n)\}} \quad (2.2)$$

onde $a(n)$ é a média das distâncias da n-ésima observação a todas as outras dentro do mesmo *cluster*, e $b(n)$ é a média das distâncias dessa n-ésima observação a todas as outras alocadas no *cluster* mais próximo. Por ser uma medida baseada apenas em distâncias, o SI independe da técnica utilizada na clusterização. Ele pode ser utilizado para medir a qualidade global do procedimento de clusterização através da média de SI_n, conforme a Eq. (2.3):

$$\overline{\mathrm{SI}} = \frac{\sum_{n=1}^{N} \mathrm{SI}_n}{N} \quad (2.3)$$

onde n representa o índice da observação e N representa o total de observações.

Em termos de procedimentos de seleção de variáveis, diversos estudos aprofundam a questão da atribuição de pesos às variáveis de clusterização. Friedman e Meulman (2004)

atribuem pesos para identificar o subgrupo de variáveis mais relevantes para clusterização, abrindo a possibilidade de atribuição de pesos diferentes a uma mesma variável que faça parte de mais de um subgrupo, seguindo critérios subjetivos. Huang *et al*. (2005) apresentam uma abordagem automática de atribuição de pesos para utilização com clusterização do tipo *k-means*; os autores realizam a avaliação das responsabilidades relativas das variáveis a cada iteração do algoritmo.

Baseados em abordagens distintas, Gnanadesikan *et al*. (1995) observam que selecionar variáveis é melhor do que atribuir-lhes pesos. Li *et al*. (2008) mostram como variáveis irrelevantes, mesmo com pesos baixos, prejudicam a clusterização, distorcendo resultados. Brusco (2004) afirma que a grande vantagem da seleção de variáveis, frente à atribuição de pesos às mesmas, é que o efeito indesejável das variáveis que poderiam mascarar a definição das estruturas dos *clusters* é totalmente excluído. Raftery e Dean (2006) realizam a seleção das variáveis dividindo-as em dois grupos complementares; o grupo das variáveis relevantes e o das irrelevantes, relacionando-as através de regressão linear. Bouveyron *et al*. (2007) realizam a seleção de variáveis após a redução da dimensão dos dados, estudando problemas de grandes dimensões. Ainda, Steinley e Brusco (2008a) realizam a comparação de diferentes métodos de seleção de variáveis para clusterização, apontando o ganho resultante dos métodos baseados no uso de informação referente à variância dos dados.

2.2.2 Análise de componentes principais

De acordo com Anderson (2003), componentes principais são combinações lineares das variáveis originais, transformando-as em um novo sistema de coordenadas ortogonais. Para Hair *et al*. (1995), a ACP é um método de extração dos fatores da análise fatorial, sendo indicado para avaliar-se a variância total do sistema em análise. Tal variância é composta pela variância comum (compartilhada entre todas as variáveis), variância específica (creditada a uma única variável) e variância do erro (devida ao processo de obtenção dos dados, erros de medidas ou componentes aleatórios no fenômeno medido). Complementarmente, Jolliffe (2002) considera a ACP como um método de redução de dimensionalidade de um conjunto de dados, explicando a maior parte da variabilidade do sistema. Os fundamentos matemáticos da ACP são agora apresentados.

Considere **x** um vetor de P variáveis. O primeiro componente principal é definido como $\alpha_1^T\mathbf{x}$, tal que os elementos de **x** tenham máxima variância, onde $\alpha_1^T = [\alpha_{11}\ \alpha_{12}\ ...\ \alpha_{1p}]$. O

segundo componente é definido como $\alpha_2^T\mathbf{x}$, não correlacionado com $\alpha_1^T\mathbf{x}$, e com os elementos de **x** tendo a máxima variância possível. Os vetores α_j são autovetores da matriz **Σ**, tida como a matriz de variâcias e covariâncias de **x**. Por fim, impõe-se à formulação de maximização de variância entre os componentes a restrição $\alpha_j^T\alpha_j = 1$, forçando o comprimento unitário nos autovetores. Nessa notação, cada autovetor α_j está relacionado com λ_j, o *j*-ésimo maior autovalor da matriz **Σ**. O problema resume-se a maximizar a variância de $\alpha_1^T\mathbf{x} = \alpha_1^T\mathbf{\Sigma}\alpha_1$, sujeito à restrição $\alpha_1^T\alpha_1 = 1$. A abordagem mais usual é utilizar a técnica dos multiplicadores de Lagrange, como na Eq. (2.4).

$$\text{Max: } \alpha_1^T\mathbf{\Sigma}\,\alpha_1 - \upsilon\,(\alpha_1^T\,\alpha_1 - 1) \tag{2.4}$$

onde υ é um multiplicador de Lagrange. O raciocínio pode ser estendido para a obtenção das demais componentes.

A ACP fornece três argumentos relevantes para sua análise: (*i*) os componentes principais (autovetores de **Σ**), representados pelos coeficientes (pesos) das variáveis em cada um dos componentes gerados; (*ii*) os dados representados no espaço dos componentes principais; e (*iii*) os autovalores de **Σ**. Cada componente principal responde por uma parcela da variância total dos dados. O valor da variância que um componente α_k representa é igual ao autovalor λ_k relacionado a esse componente. Assim, o percentual de variância retido por um componente é igual ao autovalor a ele relacionado dividido pela soma de todos os autovalores de **Σ**.

Diversos estudos têm integrado ACP em procedimentos de clusterização. Ding e He (2004) atestam a relação existente entre ACP e *k-means*, afirmando que os componentes principais gerados pela ACP são a solução contínua para os indicadores de grupos gerados na clusterização pelo método *k-means*. Honda *et al*. (2009) utilizam ACP para auxiliar no cálculo desses mesmos indicadores, aumentando o peso de variáveis que mais contribuem para a definição das estruturas dos agrupamentos. Urtubia *et al*. (2007) utilizam ACP para redução da dimensão do problema de classificação da fermentação de vinhos, utilizando informação dos três primeiros componentes principais, enquanto que Filipovych *et al*. (2011) aplicam ACP em imagens cerebrais obtidas por ressonância magnética, utilizando a informação retida pelos componentes principais para agrupar regiões do cérebro de acordo com sua capacidade de separação. Por fim, Yücel e Sultanoğlu (2012) estudam a presença de elementos químicos na composição de amostras de mel provenientes de diferentes localidades; a clusterização no

espaço gerado pelos dois primeiros componentes indicou a formação de três grupos distintos, sendo um deles composto por amostras provenientes de regiões com alta industrialização.

2.3 Método

A abordagem proposta utiliza uma ferramenta multivariada para identificar as variáveis mais relevantes para a clusterização. Segundo Steinley e Brusco (2008b) e Anzanello e Fogliatto (2011), assume-se que maior variância sugere que a variável seja mais dispersa e, com isso, tenha mais capacidade de diferenciar (e então inserir) observações em grupos, quando comparada a variáveis com menores variâncias. Para tanto, utiliza-se a ACP para gerar um índice de importância de variáveis; a informação contida nas variáveis latentes (ou seja, os pesos atribuídos às variáveis nos componentes principais) é utilizada para determinar a importância das variáveis originais.

O método proposto para seleção de variáveis para clusterização é operacionalizado em 6 passos: (1) Aplicar a ACP nos dados. (2) Definir um índice de importância das variáveis, utilizando as informações fornecidas pela ACP. (3) Definir um número limite de *clusters* a serem formados. (4) Para cada número de *clusters* em (3), incluir sistematicamente as variáveis apontadas como mais relevantes pelo índice de acordo com uma sistemática do tipo *forward*, realizar a clusterização utilizando as variáveis selecionadas e avaliar seu desempenho através do *Silhouette Index* (SI). (5) Retornar ao passo (4), alterando o número de *clusters*. (6) Identificar o número de *clusters* e as variáveis que conduzem ao SI máximo. Esses passos são detalhados na sequência.

2.3.1 Passo 1 – Aplicação da ACP

Inicialmente, deve-se preparar o conjunto de dados para análise. Tendo em vista que usualmente deseja-se obter agrupamentos de observações descritas por variáveis de diversas magnitudes, recomenda-se a normalização dos dados para garantir a consistência da clusterização (a qual é afetada pela magnitude das escalas das variáveis no cálculo das distâncias entre observações).

Na sequência, aplica-se a ACP nos dados. Dentre os parâmetros gerados pela análise, são de interesse para o estudo proposto os coeficientes das variáveis nos componentes (pesos), e os autovalores de $\mathbf{\Sigma}$.

2.3.2 Passo 2 – Geração do índice de importância das variáveis

Nesta etapa, calcula-se o Índice de Importância da Variável, IIV_p, para as P variáveis. Esse índice leva em consideração o peso α_{jp} da variável em cada um dos j componentes principais e a variância explicada por cada um desses j componentes (autovalores λ_j), conforme a Eq. (2.5). Quanto maior o valor do índice, mais importante é considerada a variável para explicação da variabilidade nos dados.

$$IIV_p = \sum_{j=1}^{J} |\alpha_{jp}| . \lambda_j \tag{2.5}$$

A princípio, o índice foi definido para levar em consideração a variância explicada por todos os componentes ($j = P$). Contudo, pode-se definir o índice utilizando apenas os componentes com maior capacidade de explicação ($j < P$). Pode-se, também, definir um valor percentual para a variância total explicada, utilizando os j's componentes que atinjam esse valor. Nesse estudo, definiu-se que serão utilizadas para a composição do IIV_p os J primeiros componentes principais responsáveis por 90% da variância dos dados.

Uma vez calculado o *IIV* para as variáveis, as mesmas são ordenadas de forma decrescente em função do valor do índice.

2.3.3 Passo 3 – Definição do intervalo de variação do número de *clusters* (*k*)

A escolha do intervalo de variação de *k*, número de *clusters*, é um passo importante do método. É bastante lógico que devam existir ao menos 2 grupos distintos entre as observações, caso contrário, considera-se não haver diferenças significativas entre essas observações. Assim, o limite inferior do intervalo de variação deverá ser 2; a definição de um limite superior, contudo, não é trivial, sendo definido por especialistas de acordo com conhecimento do sistema que está sendo agrupado.

2.3.4 Passo 4 – Inclusão das variáveis relevantes, clusterização das observações e avaliação do SI

O procedimento adotado para a clusterização é não-hierárquico, do tipo *k-means*. O subconjunto inicial de variáveis a ser testado parte de duas variáveis (as duas com os maiores valores de *IIV*), visto que agrupamentos com base em uma única variável podem apresentar comportamento instável (Anzanello *et al*., 2009). Concluída a clusterização, avalia-se a qualidade do agrupamento gerado através do SI médio de todas as observações, conforme a Eq. (2.3).

Na sequência, a terceira variável tida como mais importante pelo *IIV* é inserida no subconjunto de variáveis, uma nova clusterização é executada, e o valor do SI médio das

observações é armazenado. Tal procedimento iterativo é mantido até que todas as *P* variáveis disponíveis sejam inseridas no subconjunto de variáveis utilizadas para a clusterização. Concluída a clusterização com todas as variáveis, tem-se o valor do SI médio para cada par ordenado (*m*,*k*), no qual *m* representa o número de variáveis utilizadas na clusterização e *k*, o número de grupos formados.

2.3.5 Passo 5 – Fazer $k = k + 1$ e retornar ao Passo 4

Para determinar o número de *clusters* mais adequado, define-se $k = k + 1$ e recomeça-se o procedimento de inserção de variáveis, clusterização de observações e avaliação da qualidade dos agrupamentos via SI. Esse procedimento é repetido até que o limite superior em *k* seja atingido. Neste estudo, investigou-se a separação em até 5 *clusters*.

2.3.6 Passo 6 – Identificar o melhor número de *clusters* e as variáveis para clusterização

O valor máximo do SI médio para o intervalo de número de *clusters* testado é identificado; tal valor indica o melhor número de *clusters*, assim como as variáveis recomendadas pelo método. Alternativamente, pode-se definir o melhor conjunto de variáveis para determinado número de *clusters* a serem formados.

2.4 Exemplos numéricos

O método proposto foi aplicado em três bancos de dados. O primeiro banco, CURVAS, foi utilizado por Anzanello e Fogliatto (2011); as variáveis de clusterização são parâmetros oriundos de modelagem de curva de aprendizado que caracterizam desempenho de trabalhadores, e as observações referem-se a trabalhadores a serem agrupados de acordo com seus perfis de adaptação a procedimentos repetitivos. Optou-se por utilizar o banco CURVAS para que pudessem ser feitas comparações entre os desempenhos dos métodos. Esse banco consiste em 20 observações (trabalhadores) descritas por 12 variáveis. Os outros bancos foram escolhidos para verificar a capacidade do método, não sendo encontrada nenhuma utilização dos mesmos em problemas de clusterização. Um desses bancos contém informações sobre pacientes com câncer de mama (BCANCER), consistindo em 569 pacientes (observações) avaliadas através de 30 variáveis (disponível em http://archive.ics.uci.edu/ml/datasets.html). O terceiro banco aborda movimentos de mão na Língua Brasileira de Sinais (LIBRAS), com 360 observações descritas por 90 variáveis (disponível em http://archive.ics.uci.edu/ml/datasets.html).

O método proposto foi implementado utilizando-se o aplicativo MATLAB®, versão 7.0.0.19920 (R14). Os tempos de processamento computados incluem a leitura do banco de

dados e a aplicação dos passos descritos no método. Para o banco CURVAS, o tempo foi inferior a 1 segundo; para o banco BCANCER o tempo de processamento foi de 30 segundos, aproximadamente; para o banco LIBRAS o tempo foi de 1 minuto e 26 segundos, aproximadamente.

As duas primeiras colunas da Tabela 2.1 apresentam o índice de importância, *IIV*, para cada variável e a identificação dessa variável nos dados originais (CURVAS). As duas colunas restantes mostram o percentual de variância explicada por cada um dos doze componentes principais e o percentual cumulativo, utilizado para identificar o ponto de corte no cálculo do IIV_p; para o banco CURVAS, utilizaram-se os pesos das variáveis nos seis primeiros componentes para obtenção do índice (explicando mais de 90% da variância do sistema).

Definiu-se 5 como o número máximo de *clusters* a serem testados. As variáveis ordenadas foram então inseridas uma a uma para realizar a clusterização, conforme as Seções 2.3.4 e 2.3.5. Os valores do SI médio gerados pela inserção das variáveis no procedimento de clusterização para o banco CURVAS podem ser observados na Tabela 2.2.

Percebe-se que duas variáveis (12 e 2, identificadas da Tabela 2.1) devem ser retidas para a formação de cinco grupos distintos, gerando um SI médio de 0,65. Tal valor representa um acréscimo de 62,5% na qualidade dos agrupamentos formados em comparação à utilização de todas as variáveis; para este resultado, utilizaram-se apenas 17% das variáveis disponíveis (2 de um total de 12).

Tabela 2.1 – Valor do *IIV*, identificação da variável e variâncias explicadas pelos componentes principais – Banco CURVAS

Variáveis Originais		Componentes Principais	
IIV_j	Identificação da variável	Variância Explicada (%)	Variância Cumulativa (%)
1,8121	12	32,40	32,40
1,6434	2	24,47	56,87
1,6370	8	15,04	71,91
1,6268	7	7,49	79,40
1,5817	11	6,15	85,55
1,5196	9	4,49	90,04
1,1911	4	4,26	94,30
0,9823	5	2,60	96,90
0,7769	6	2,05	98,95
0,5512	1	0,63	99,58
0,5438	3	0,42	100,00
0,3060	10	0,00	100,00

Apesar do aumento de qualidade na clusterização, com reduzido percentual de variáveis retidas, nenhum dos resultados obtidos pelo método proposto superou os valores obtidos por Anzanello e Fogliatto (2011), onde o valor médio de SI superou 0,90.

Tabela 2.2 – SI médio – Banco CURVAS

Número de variáveis utilizadas na clusterização	Número de *clusters* (k)			
	2	3	4	5
2	**0,60**	**0,63**	**0,60**	**0,65**
3	0,43	0,53	0,55	0,61
4	0,56	0,51	0,51	0,54
5	0,54	0,45	0,50	0,51
6	0,49	0,41	0,41	0,50
12	0,23	0,34	0,31	0,40

A Tabela 2.3 apresenta resultados relativos ao banco de dados BCANCER, onde foram utilizados os seis primeiros componentes principais no cálculo do IIV_p. A melhor escolha para fins de clusterização é selecionar duas variáveis (28 e 8) para a formação de dois *clusters*. Tal seleção eleva a qualidade do agrupamento de 0,60 (com todas as variáveis) para 0,79 (com as variáveis selecionadas). Em termos percentuais, obteve-se uma melhora média na qualidade dos agrupamentos da ordem de 70%, quando considerados os agrupamentos em 2 (melhora de 32%), 3 (melhora de 39%), 4 (melhora de 85%) e 5 *clusters* (melhora de 130%), retendo em média menos de 8% das variáveis originais (2 variáveis em 30 para 2, 4 e 5 *clusters*, e 3 variáveis para 3 *clusters*).

Tabela 2.3 – SI médio – Banco BCANCER

Número de variáveis utilizadas na clusterização	Número de *clusters* (k)			
	2	3	4	5
2	**0,79**	0,66	**0,63**	**0,63**
3	0,78	**0,68**	0,63	0,57
4	0,76	0,65	0,57	0,51
5	0,74	0,63	0,57	0,47
6	0,73	0,63	0,52	0,51
30	0,60	0,49	0,34	0,27

A Figura 2.1 traz uma representação gráfica do SI para o banco BCANCER, onde cada linha horizontal representa o SI de uma observação. Percebem-se algumas observações alocadas ao grupo errado, evidenciadas pelos valores negativos do SI.

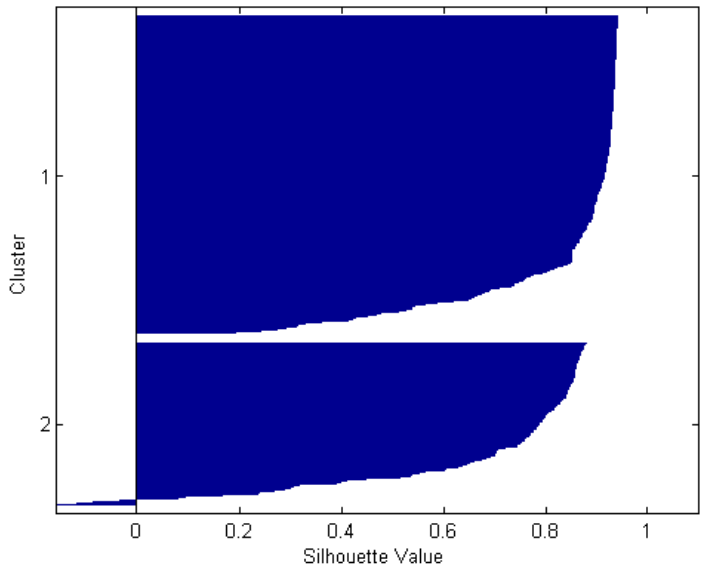

Figura 2.1 – Gráfico do SI para 2 *clusters* utilizando 2 variáveis (28 e 8) – BCANCER

A Figura 2.2 mostra a separação no mesmo banco de dados para 2 *clusters* utilizando todas as variáveis. É visível o maior número de alocações equivocadas, bem como os menores valores para o SI.

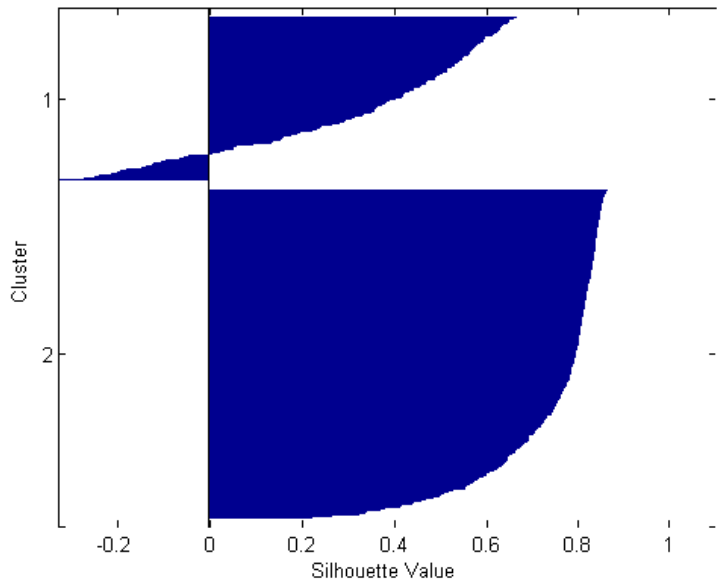

Figura 2.2 – Gráfico do SI para 2 *clusters* utilizando todas as variáveis – BCANCER

Comparativamente à utilização de todas as variáveis, percebe-se que o método proposto se mostrou satisfatório por dois aspectos principais: (*i*) reduziu o percentual de alocações erradas das observações nos *clusters*, conforme se observa na menor quantidade de valores negativos de SI na Figura 2.1; e (*ii*) os dois *clusters* formados têm grande número de

observações com valores elevados de SI, da ordem de 0,80, refletindo que os agrupamentos formados são distintos entre si e mais homogêneos do que os obtidos quando se utilizam todas as variáveis.

A Tabela 2.4 apresenta os resultados para o banco LIBRAS. Apesar de contar com 90 variáveis, apenas seis componentes principais foram utilizados no cálculo do IIV_p, uma vez que explicam 90,70% da variância total dos dados.

Tabela 2.4 – SI médio – Banco LIBRAS

Número de variáveis utilizadas na clusterização	Número de *clusters* (k)			
	2	3	4	5
2	**0,74**	**0,72**	**0,66**	**0,70**
3	0,73	0,69	0,66	0,65
4	0,73	0,68	0,61	0,62
5	0,73	0,69	0,61	0,63
6	0,71	0,66	0,59	0,57
90	0,34	0,29	0,30	0,32

O melhor resultado de clusterização se dá com duas variáveis (89 e 85) e 2 *clusters*. Esse conjunto elevou o SI médio de 0,34 (utilizando todas as variáveis do banco) para 0,74 (utilizando as variáveis selecionadas). Isto representou um ganho de 117% no SI médio para a formação de 2 *clusters*. Para 3, 4 e 5 *clusters*, os ganhos foram respectivamente de 148%, 120% e 118%, utilizando apenas as 2 variáveis selecionadas (2% das variáveis disponíveis).

Com base nos resultados acima, comprova-se que é preferível utilizar apenas um subconjunto das variáveis para geração de agrupamentos. Para todos os dados testados, os valores do SI médio obtidos com as variáveis selecionadas pelo método são consideravelmente maiores do que quando toda a informação disponível é utilizada. Mesmo que os resultados não tenham superado outros métodos já existentes, demonstra-se que o método proposto tem potencial de utilização como ferramenta de auxílio na escolha das melhores variáveis para clusterização, uma vez que o *IIV* constitui-se em um critério simples.

2.5 Conclusões

As técnicas de clusterização são extremamente importantes em diversos cenários práticos, incluindo aplicações industriais, médicas e sociais. Este artigo propôs uma sistemática de seleção de variáveis para clusterização, baseada nos parâmetros oriundos da ACP. A metodologia consiste em: (1) Aplicar a ACP nos dados. (2) Gerar o IIV_p. (3) Definir o número limite k de *clusters* a serem formados. (4) Para cada k em (3), inserir as variáveis mais relevantes no subconjunto de variáveis de clusterização, realizar a clusterização e avaliar

seu desempenho através do SI. (5) Retornar ao passo (4), fazendo $k = k + 1$. (6) Identificar o número de *clusters* e as variáveis que conduzem ao SI máximo.

Dentre as proposições deste artigo, um novo índice de importância de variáveis baseado em ACP foi sugerido, garantindo que informação relativa à proporção da variância total relacionada a cada variável seja utilizada na seleção das mais importantes para clusterização. O método proposto foi aplicado a três bancos de dados de diferentes contextos (produção industrial, saúde e linguagem), retendo menos de 10% das variáveis originais, em média, e obtendo resultados com ganho médio de 100% na qualidade quando comparado à utilização de todas as variáveis originais.

Trabalhos futuros incluem estudos no sentido de, uma vez obtido o ordenamento através do IIV_p, executar o método proposto por Anzanello e Fogliatto (2011) para um subconjunto das variáveis ordenadas. Outra possibilidade seria executar, para esse subconjunto definido, um método exaustivo de busca pelas melhores variáveis. Essa opção, contudo, deve ser estudada com atenção, uma vez que, para um universo de variáveis da ordem de milhares, procedimentos exaustivos podem inviabilizar a enumeração total.

2.6 Referências

Agard, B., Penz, B., 2009. A simulated annealing method based on a clustering approach to determine bills of materials for a large product family. **International Journal of Production Economics** 117 (2), 389-401.

Anderson, T.W., 2003. **An Introduction to Multivariate Statistical Analysis** third ed. John Wiley & Sons, Inc. Hoboken, New Jersey.

Anzanello, M.J., Albin, S.L., Chaovalitwongse, W., 2009. Selecting the Best variables for classifying production batches into two quality levels. **Chemometrics and Intelligent Laboratories Systems** 97 (2), 111-117.

Anzanello, M.J., Fogliatto, F.S., 2011. Selecting the Best clustering variables for grouping mass-customized products involving workers' learning**. International Journal of Production Economics** 130 (2), 268-276.

Bouveyron, C., Girard, S., Schmid, C., 2007. High-dimensional data clustering. **Computational Statistics and Data Analysis** 52, 502-519.

Brusco, M.J., 2004. Clustering binary data in the presence of masking variables. **Psychological Methods**, 9, 510-523.

Brusco, M.J., Cradit, J.D., 2001. A variable-selection heuristic for k-means clustering. **Psychometrika** 66 (2), 249-270.

Brusco, M.J., Steinley, D., Cradit, J.D., Singh, R., 2012. Emergent clustering methods for empirical OM research. **Journal of Operations Management** 30 (6), 454-466.

Ding, C., He, X., 2004. *K*-means clustering via principal component analysis. In: Russ Greiner, Dale Schuurmans (Eds**.) Proceedings of the 21st International Machine Learning Conference**, ACM Press, 225-232.

Duda, R.O., Hart, P.E., Stork, D.G., 2001**. Pattern classification** second edition John Wiley & Sons, New York.

Filipovych, R., Resnick, S.M., Davatzikos, C., 2011. Semi-supervised cluster analysis of imaging data. **Neuroimage** 54 (3), 2185-2197.

Fowlkes, E., Gnanadesikan, R., Kettenring, J., 1988. Variable selection in clustering. **Journal of Classification** 5 (2), 205-228.

Friedman, J.H., Meulman, J.J., 2004. Clustering objects on subsets of attributes (with discussion). **Journal of the Royal Statistical Society**, Series B 66, 815-849.

Gnanadesikan, R., Kettenring, J., Tsao, S., 1995. Weighting and selection of variables for cluster analysis. **Journal of Classification** 12 (1), 113-136.

Hair, J., Anderson, R., Tatham, R., Black, W., 1995. **Multivariate Data Analysis with Readings** fourth ed. Prentice-Hall Inc., New Jersey.

Honda, K., Notsu, A., Ichihashi, K., 2009. PCA-guided k-means with variable weighting and its application to document clustering. **Lecture Notes in Computer Science**, 5861, 282-292.

Huang, J.Z., Ng, M.K., Rong, H. Li, Z., 2005. Automated variable weighting in k-means type clustering. **IEEE Transactions on Pattern Analysis and Machine Intelligence** 27 (5),657-668.

Jolliffe, I.T., 2002. **Principal Component Analysis** second ed. Springer-Verlag New York.

Kaufman, L., Rousseeuw, P., 2005. **Finding Groups in Data: an Introduction to Cluster Analysis.** Wiley Interscience, New Jersey.

Li, Y., Dong, M., Hua, J., 2008. Localized feature selection for clustering. **Pattern Recognition Letters** 29 (1), 10-18.

Maugis, C., Celeux, G., Martin-Magniette, M., 2009. Variable selection for clustering with Gaussian mixture models. **Biometrics** 65 (3), 701-709.

Raftery, A.E., Dean, N., 2006. Variable selection for model-based clustering. **Journal of the American Statistical Association** 101, 168-178.

Steinley, D., Brusco, M.J., 2008a. Selection of variables in cluster analysis: an empirical comparison of eight procedures. **Psychometrika** 73 (1), 125-144.

Steinley, D., Brusco, M.J., 2008b. A new variable weighting and selection procedure for K-means cluster analysis. **Multivariate Behavioral Research** 43 (1), 77-108.

Urtubia, A., Perrez-Correa, J., Soto, A., Pszczolkowski, P., 2007. Using data mining techniques to predict industrial wine problem fermentation. **Food Control** 18 (1), 1512-1517.

Yücel, Y., Sultanoğlu, P., 2012. Determination of industrial pollution effects on citrus honeys with chemometric approach. **Food Chemistry** 135, 170-178.

3 Segundo Artigo: Avaliação da robustez de uma sistemática de seleção de variáveis para clusterização através de experimentos de simulação

Victor Leonardo Cervo

Michel Jose Anzanello

Artigo a ser enviado para publicação na revista Gestão e Produção

Resumo

A separação de observações em grupos distintos, com alto grau de semelhança entre seus elementos, é uma tarefa de extrema importância para várias áreas do conhecimento. Metodologias de seleção de variáveis em aplicações de clusterização visam melhorar a qualidade dos agrupamentos gerados, através da exclusão de variáveis ruidosas e irrelevantes que descrevem as observações. Neste artigo, a metodologia proposta em Cervo e Anzanello (2013) é avaliada através de simulação, a fim de medir o desempenho do método em contextos com distintos níveis de correlação e ruído nas variáveis, bem como distintos tamanhos de amostra. Os resultados indicam que o método é satisfatoriamente robusto para redução do número de variáveis em procedimentos de clusterização.

Palavras-chaves: análise de clusterização, seleção de variáveis, simulação.

Robustness' assessment of a framework for clustering variable selection through simulation experiments

Abstract

The separation of observations into distinct groups with a high degree of similarity between their elements is an extremely important task in many fields of knowledge. Variable selection methods aim to improve *clusters* quality by removing noisy and irrelevant variables. In this work, the robustness of Cervo and Anzanello (2013)'s propositions is evaluated through simulation, assessing the method's performance on different levels of variable noise

and correlation, and number of observations. Results indicate that the method is satisfactorily robust to clustering variable selection.

Keywords: clustering analysis, variable selection, simulation

3.1 Introdução

Técnicas de clusterização buscam definir agrupamentos (*clusters*) em um conjunto de observações tipicamente descritas por variáveis. Os grupos formados devem conter elementos com grande similaridade entre si (Hair *et al*., 1995), e forte diferenciação em relação a observações inseridas em outros *clusters* (Agard e Penz, 2009). Frey e Dueck (2007) investigam a obtenção de *clusters* através de uma sistemática de avaliação da relação entre as observações, escolhendo algumas delas como centro de *clusters*, ou exemplares.

Uma vez realizada a clusterização, a qualidade do procedimento pode ser avaliada através de uma métrica como o *Silhouette Index* (SI). Essa medida avalia o quanto uma observação é semelhante às demais que fazem parte do mesmo *cluster*, comparado com as observações alocadas ao *cluster* vizinho (Kaufmann e Rousseeuw, 2005). Calculando-se a média de todos os SI das observações alocadas a *clusters*, têm-se uma medida geral da qualidade dos agrupamentos gerados e, consequentemente, do desempenho do procedimento de clusterização (TABOADA; COIT, 2008).

Um grande desafio para a análise de clusterização é a escolha de variáveis relevantes e detentoras de informações diferenciadoras que contribuam para a identificação de *clusters* adequados (Milligan, 1989; Brusco e Cradit, 2001; Steinley e Brusco, 2008a; Maugis *et al*., 2009). Em abordagens típicas, esses subconjuntos podem ser obtidos através da atribuição de pesos às variáveis de clusterização, indicando sua relevância na formação dos agrupamentos. Alternativamente, diversos autores apontam as vantagens de selecionar variáveis (ou seja, atribuir pesos zero ou um às mesmas), uma vez que a seleção exclui efeitos de mascaramento na identificação da real estrutura dos *clusters* (Brusco, 2004). Milligan (1980) e Li *et al*. (2008) corroboram essa informação, apontando que a atribuição de pesos baixos às variáveis irrelevantes pode distorcer significativamente a qualidade dos agrupamentos formados. A fim de identificar as variáveis mais relevantes para classificação e clusterização, diversos autores apoiam-se em ferramentas multivariadas: Anzanello *et al*. (2009) utilizam os parâmetros da regressão PLS para gerar índices de importância para as variáveis em procedimentos de classificação usando a técnica de classificação *k*-Nearest Neighbor, enquanto Fowlkes *et al*. (1988) e Gnanadesikan *et al*. (1995) aplicam ANOVA multivariada para estabelecer um

critério de separação para as variáveis selecionadas. Outros autores usam informação relativa à variância das variáveis para realizar a seleção das mesmas: Steinley e Brusco (2008b) mostram que a parcela de variância contida por uma variável está diretamente relacionada à importância desta variável na recuperação dos *clusters*. Com propósitos semelhantes, Cervo e Anzanello (2013) geram um índice de importância a partir dos parâmetros oriundos da Análise de Componentes Principais (ACP), o qual guia a inserção das variáveis mais relevantes em uma sistemática de clusterização. Embora tal sistemática garanta que as variáveis com maiores pesos nos primeiros componentes principais obtidos sejam utilizadas na geração de agrupamentos, não existem maiores conclusões acerca de sua robustez frente a distintos níveis de correlação e ruído nas variáveis descritivas das observações.

Este artigo avalia a robustez do método proposto por Cervo e Anzanello (2013) valendo-se de experimentos de simulação. Para tanto, são gerados 108 experimentos baseados em distintos níveis de correlação e ruído das variáveis, além do número de observações a serem clusterizadas.

O artigo está organizado conforme segue. A Seção 3.2 apresenta uma breve fundamentação teórica sobre seleção de variáveis para clusterização. A Seção 3.3 apresenta o projeto da simulação dos dados e o método utilizado para clusterização. Os resultados obtidos e a discussão acerca dos mesmos são apresentados na Seção 3.4. A Seção 3.5 traz as conclusões e os direcionamentos futuros, encerrando o artigo.

3.2 Fundamentação teórica

3.2.1 Seleção de variáveis para clusterização

A clusterização é uma técnica de análise multivariada que tem por objetivo fundamental alocar observações em *clusters*, de forma que as similaridades entre observações dentro de um mesmo *cluster* sejam elevadas, enquanto os *clusters* devem ser distintos entre si (Hair *et al*., 1995). Existem, de forma geral, 2 tipos de métodos para realizar a clusterização: (1) métodos hierárquicos e (2) métodos não hierárquicos. Os métodos hierárquicos definem uma estrutura de hierarquia entre as observações; esta hierarquia comumente é visualizada através de um dendograma. Este método é indicado para a avaliação quanto ao número de *clusters* que devem ser gerados, uma vez que permite, visualmente, identificar o número de *clusters* adequado, sendo este definido através de cortes nos ramos do dendograma (Hair *et al*., 1995). A classe de métodos não hierárquicos, da qual faz parte o *k-means*, não define hierarquia entre observações. Esse tipo de método é chamado de método por particionamento,

uma vez que implementa a divisão em *k clusters*, número este que deve ser definido no momento da implementação. É usual, em trabalhos envolvendo este tipo de método, realizar a clusterização para um intervalo de *k*'s, a fim de avaliar o número k_c ideal de *clusters* (ANZANELLO; FOGLIATTO, 2011).

Estudos mostram que a recuperação de uma estrutura de *clusters* depende de um subconjunto reduzido de variáveis. A inserção de variáveis inadequadas nesse subconjunto pode ser inútil, ou até mesmo impedir a correta geração dos *clusters* de observações. Vários estudos indicam os problemas e a degradação da qualidade da clusterização decorrente da utilização de variáveis incapazes de diferenciar observações em grupos (MILLIGAN, 1980; LI *et al*., 2008; MAUGIS *et al*., 2009).

Ferramentas de seleção de variáveis podem ser entendidas como classes especiais de problemas de atribuição de pesos às variáveis de clusterização, com a atribuição de peso zero às variáveis ruidosas e peso unitário às variáveis que se acredita serem relevantes para a clusterização (Brusco e Cradit, 2001). Brusco (2004) afirma que a seleção de variáveis conduz a melhores resultados para clusterização, uma vez que elimina o efeito das variáveis que não definem a estrutura de *clusters*. Com o objetivo de obter segmentos (grupos) de mercado representativos, Liu e Ong (2008) utilizam um algoritmo genético para selecionar as melhores variáveis a serem utilizadas na clusterização por *k-means*, enquanto que Karimi e Hemmateenejad (2013) aplicam uma metodologia de seleção de variáveis para clusterização de pacientes de câncer ovariano e de próstata, com bons resultados para classificação de pacientes. Poon *et al*. (2013) abordam a questão de seleção de variáveis em bancos de dados com elevados volumes de informação e apresentam a ideia de determinação de facetas, ou visões diferenciadas do conjunto de dados, baseadas em algumas variáveis; segundo os autores, uma metodologia de determinação das facetas produziu melhores resultados do que a seleção de variáveis executada por procedimentos usuais.

Diversos autores têm projetado experimentos de simulação a fim de estabelecer resultados consistentes para técnicas de clusterização, incluindo classificação e segmentação de mercado. Bellec *et al*. (2010) utilizam *k-means* em várias amostras de dados obtidas por *bootstrapping* a partir de imagens de ressonância magnética; os autores identificam as variáveis mais estáveis que melhor definem redes neurais de interesse. Andrews *et al*. (2010) comparam métodos baseados em modelos (como o método de mistura finitas, que assume distribuições características para as variáveis) e métodos não baseados em modelos (como o *k-means*), que podem ser utilizados para obtenção de partições de múltiplas bases de clientes

na segmentação de mercado; resultados obtidos pela simulação indicam que os métodos que consideram modelos estatísticos têm maior capacidade de recuperar a estrutura de segmentos (*clusters*). Dean e Raftery (2010) utilizam técnicas de simulação para avaliar a sistemática de seleção de variáveis para clusterização de dados categóricos; os autores consideram 2 modelos distintos para decidir se uma variável será ou não incluída no subconjunto de variáveis de clusterização. Ainda, Andrews e McNicholas (2012) utilizam dados simulados para avaliar a capacidade de classificação de um método baseado em mistura de modelos que usa distribuição *t* multivariada para caracterizar os componentes, observando bons resultados.

3.2.2 Análise de componentes principais – ACP

A ACP consiste em representar um conjunto de dados qualquer, com observações descritas por variáveis, em um novo sistema de coordenadas, ortogonais entre si, com número de eixos igual ao número original de variáveis (Anderson, 2003). Esses eixos ortogonais são os componentes principais, e estão diretamente relacionados aos autovetores e autovalores da matriz de covariância dos dados. Cervo e Anzanello (2013) apresentam os tópicos de interesse para a implementação da sistemática proposta, enquanto Jolliffe (2002) e Anderson (2003) apresentam os fundamentos matemáticos da técnica.

Os parâmetros derivados da ACP utilizados por Cervo e Anzanello (2013) são os pesos de cada variável em cada componente principal e os autovalores da matriz de covariância dos dados. Esses parâmetros são combinados para gerar um índice de importância de variável (IIV_p), o qual guiará o processo de inclusão de variáveis no processo de clusterização.

3.3 Método

A abordagem proposta gera diferentes cenários de dados simulados para avaliar o desempenho da sistemática de seleção proposta por Cervo e Anzanello (2013). São geradas variáveis descritivas de observações a serem agrupadas de acordo com 3 fatores, cada qual com 3 níveis: ruído entre as variáveis, correlação entre as variáveis, e proporção de observações utilizadas para gerar o IIV, todos considerados a nível alto, nominal e baixo.

3.3.1 Projeto de simulação

A simulação gerou dados baseados em dados reais de desempenho de trabalhadores submetidos a atividades manuais em uma empresa do setor calçadista. O banco de dados original consiste de 20 observações, descritas por 12 variáveis. Os bancos gerados na simulação seguem distribuições Normais Multivariadas com média $\boldsymbol{\mu}$, variâncias dadas de

acordo com uma matriz **Σ** de covariâncias e correlações de acordo com uma matriz **P**. O vetor **μ** e as matrizes **Σ** e **P** são extraídos do banco original.

Os fatores em estudo são a correlação, ruído nas variáveis e proporção das observações utilizada no cálculo do IIV. O fator correlação é investigado a 3 níveis: alto, utilizando $\mathbf{P^{1/3}}$; nominal, utilizando **P**; e baixo, utilizando $\mathbf{P^3}$. O fator ruído é investigado a 3 níveis: alto, considerando erros para mais (fazendo a média do erro igual a 1); nominal, considerando média do erro igual a 0; e baixo, considerando erros para menos (fazendo a média do erro igual a -1). O fator proporção de observações utilizadas é investigado a 3 níveis: alto, considerando a totalidade das observações disponíveis; nominal, considerando 10% das observações; baixo, considerando 1% do total de observações disponíveis. O fator proporção nos níveis médio e baixo apresenta outra condição: para o nível médio, garante-se o número mínimo de 50 observações utilizadas; para o nível baixo, garante-se o mínimo de 10 observações a serem consideradas. Essa condição visa minimizar efeitos de observações *outliers* sobre o cálculo do IIV. A Tabela 3.1 apresenta os fatores e os níveis do experimento.

Tabela 3.1 – Fatores e níveis do experimento

Fatores	Níveis
Correlação das variáveis	$\mathbf{P^{1/3}}$**; P;** $\mathbf{P^3}$
Média do erro	1;0;-1
Proporção de observações utilizadas	1;0,1;0,01

Para cada nível do fator correlação são gerados bancos com 100, 200, 500 e 1000 observações. Os erros também são gerados de forma a considerar esses números de observações. Os dados finais são obtidos pela soma ponderada de dados gerados a um nível do fator correlação com o fator erro, em um dos 3, conforme a Eq. (3.1). Um fator de escala faz a ponderação. Adotou-se o fator de escala igual a 10, assumindo, empiricamente, que um erro é capaz de alterar em 10% o valor esperado do dado.

$$d_{c,e} = f.d'_c + \varepsilon_e \tag{3.1}$$

onde $d_{c,e}$ representa os dados obtidos com o nível de correlação c e nível de erro e, f é o fator de escala, d'_c representa os dados brutos obtidos com nível de correlação c, ε_e representa o erro gerado com nível e.

Cada um dos 9 bancos ($d_{c,e}$) obtidos pela soma ponderada é avaliado em 3 níveis do fator proporção. Cada uma das combinações possíveis é utilizada, gerando 27 experimentos a serem analisados, com um número fixo de observações no banco.

3.3.2 Aplicação da ACP aos dados

Conforme proposto por Cervo e Anzanello (2013), os dados são inicialmente normalizados, de forma a garantir a consistência da ACP e da clusterização (Milligan e Cooper, 1988). Após a normalização, a ACP é aplicada aos dados, considerando cada um dos 3 níveis do fator proporção, e os parâmetros de interesse são coletados.

3.3.3 Geração do IIV

O cálculo do IIV_p para cada variável é baseado nos parâmetros obtidos da ACP, conforme a Eq. (3.2) (CERVO; ANZANELLO, 2013):

$$IIV_p = \sum_{j=1}^{J} |\alpha_{jp}| . \lambda_j \tag{3.2}$$

onde j identifica o j-ésimo componente principal, J identifica o componente com menor variância utilizado (J componentes que respondam por, no mínimo, 90% da variância total dos dados), λ_j é o autovalor relacionado ao j-ésimo componente principal e representa a parcela de variância representada por este componente, α_{jp} é o peso da variável p no j-ésimo componente. As variáveis são ordenadas de forma decrescente quanto ao IIV, considerando-se a que tem maior valor como mais importante.

3.3.4 Definir valores limites para o intervalo de variação de *k*

Uma vez gerado o índice, procede-se à definição do intervalo de variação do número k de *clusters*. Usualmente define-se como valor mínimo 2, tendo em vista que a técnica busca separar observações em grupos. Portanto, o valor mínimo utilizado neste artigo será $k = 2$. O valor máximo será o mesmo utilizado por Cervo e Anzanello (2013): 5. Salienta-se, contudo, que o valor máximo de k deve ser definido por especialistas com base em conhecimentos sobre o sistema estudado.

3.3.5 Realizar os procedimentos de clusterização e avaliar o SI obtido

Para cada número k de *clusters*, o procedimento de clusterização inicia com 2 variáveis selecionadas. Sugere-se um mínimo de 2 variáveis devido a instabilidades que podem ser geradas quando se considera apenas uma variável (Anzanello *et al*., 2009). Realizada a clusterização, o procedimento é avaliado pelo valor do SI médio de todas as

observações. As variáveis são sistematicamente inseridas no conjunto de variáveis de clusterização, seguindo estritamente a ordem estabelecida pelo IIV, novas clusterizações são realizadas e avaliadas pelo SI médio de todas as observações para o número *k* de *clusters*, até que todas as variáveis sejam inseridas no grupo.

3.3.6 Fazer $k = k + 1$ e retornar para 3.3.5

O desempenho da clusterização com diferentes números de *clusters* indica, dentre todas as alternativas, aquela que conduz ao maior SI médio. Nesse passo incrementa-se o valor de *k* e retoma-se a clusterização pela inclusão sistemática de variáveis, regida pelo IIV. O procedimento é repetido até ser atingido o valor máximo estabelecido para *k*.

3.3.7 Indicar o melhor número de *clusters* e as variáveis para clusterização

Identifica-se o máximo SI médio obtido; esse valor indica o número ideal de *clusters* a serem formados e as variáveis que devem ser selecionadas, ou seja, as mais indicadas para fazerem parte do grupo de variáveis de clusterização.

3.4 Resultados da simulação

A sistemática descrita foi operacionalizada conforme descrito por Cervo e Anzanello (2013): o percentual de variância explicada retida pelos componentes e utilizado no cálculo do IIV foi 90%, o número *k* de *clusters* variou de 2 a 5, e o conjunto de variáveis de clusterização contém, no mínimo, 2 variáveis.

As gerações de dados e os demais passos do método proposto foram implementados com a utilização do aplicativo MATLAB®, versão 7.0.0.19920 (R14). Os tempos de processamento computados abrangem desde a geração de dados até a obtenção dos resultados das clusterizações. Para o banco com 100 observações, o tempo aproximado foi de 35 segundos; para 200 observações, o tempo computado foi de 1 minuto e 15 segundos; para o banco com 500 0bservações, o tempo de processamento foi de 3 minutos e 55 segundos; para 1000 observações, o tempo aferido foi de 11 minutos e 40 segundos. Os resultados podem ser visualizados na Tabela 3.2, onde são apresentados o maior valor do SI médio obtido, o número *k* de *clusters* e o número *p* de variáveis retidas pelo método para cada combinação de fatores e número de observações totais dos bancos de dados.

A Figura 3.1 mostra a variação do SI médio para os 3 níveis do fator correlação, mantendo-se fixos os fatores proporção e erro, em seus níveis altos, para o banco com 500 observações.

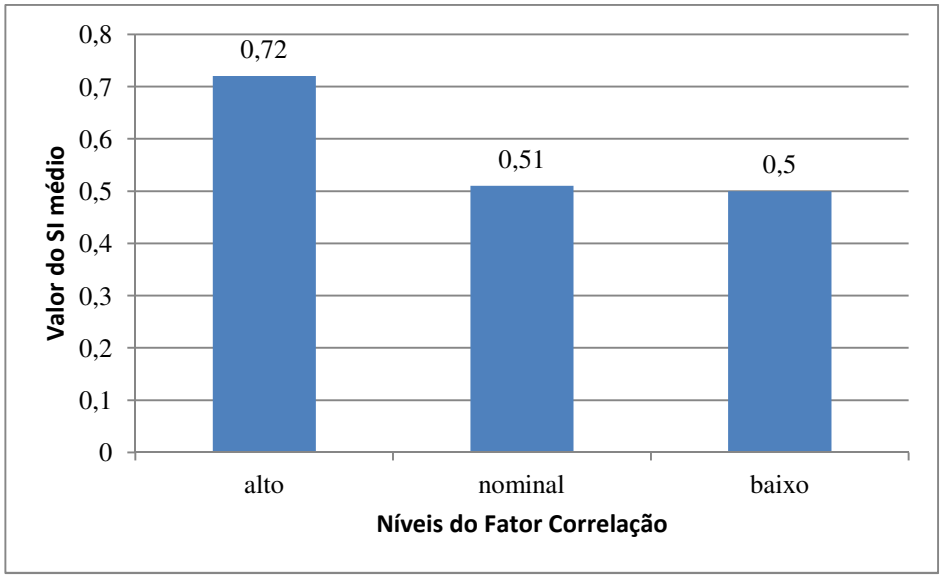

Figura 3.1 – Perfis de SI médio para o fator Correlação, a 3 níveis

Tabela 3.2 – Maiores valores de SI médio obtidos, número *k* de clusters, número *p* de variáveis selecionadas

Proporção	Correlação	Erro	100 observações			200 observações			500 observações			1000 observações		
			SI	*k*	*p*	SI	*k*	*p*	SI	*k*	*p*	SI	*k*	*p*
alto	alto	alto	0,67	2	3	0,68	2	2	0,72	2	2	0,59	2	4
		nominal	0,67	2	3	0,68	2	2	0,72	2	2	0,59	2	4
		baixo	0,67	2	3	0,68	2	2	0,72	2	2	0,59	2	4
	nominal	alto	0,54	5	2	0,51	3	2	0,51	4	2	0,53	2	2
		nominal	0,53	2	2	0,51	3	2	0,51	4	2	0,53	2	2
		baixo	0,53	2	2	0,52	3	2	0,51	4	2	0,53	2	2
	baixo	alto	0,56	3	2	0,5	5	2	0,5	4	2	0,49	3	2
		nominal	0,56	3	2	0,5	3	2	0,49	4	2	0,49	3	2
		baixo	0,56	3	2	0,5	3	2	0,49	4	2	0,49	3	2
nominal	alto	alto	0,66	2	2	0,67	2	2	0,65	2	3	0,65	2	2
		nominal	0,66	2	2	0,67	2	2	0,65	2	3	0,65	2	2
		baixo	0,66	2	2	0,67	2	2	0,65	2	3	0,65	2	2
	nominal	alto	0,53	2	2	0,51	3	2	0,52	2	2	0,51	2	2
		nominal	0,54	5	2	0,51	3	2	0,52	2	2	0,51	2	2
		baixo	0,52	5	2	0,51	3	2	0,52	2	2	0,51	2	2
	baixo	alto	0,52	3	2	0,5	2	2	0,51	3	2	0,49	3	2
		nominal	0,52	3	2	0,5	4	2	0,51	3	2	0,49	3	2
		baixo	0,52	3	2	0,5	4	2	0,5	3	2	0,49	3	2
baixo	alto	alto	0,61	2	2	0,62	2	5	0,69	2	2	0,58	2	4
		nominal	0,61	2	2	0,62	2	5	0,69	2	2	0,58	2	4
		baixo	0,61	2	2	0,62	2	5	0,69	2	2	0,58	2	4
	nominal	alto	0,56	2	2	0,51	4	2	0,53	2	3	0,59	2	2
		nominal	0,56	2	2	0,5	3	2	0,54	2	3	0,59	2	2
		baixo	0,56	2	2	0,51	4	2	0,53	2	3	0,59	2	2
	baixo	alto	0,54	3	2	0,51	3	2	0,48	3	2	0,49	3	2
		nominal	0,54	3	2	0,5	2	2	0,49	3	2	0,49	3	2
		baixo	0,55	2	2	0,5	2	2	0,49	3	2	0,49	3	2

Observa-se na Figura 3.1 que o fator correlação a nível alto conduziu o SI médio a um patamar mais elevado. Analisando-se a Tabela 3.2, constata-se o mesmo fato em quase todos os experimentos. Isto indica que o método tem seu desempenho positivamente influenciado pela maior correlação dos dados. Sugere-se que este fato seja consequência direta da aplicação da ACP: a correlação a nível alto se refletiu em aumento dos pesos das variáveis nos componentes principais mais importantes, de forma que as variáveis mais capazes de definir *clusters* tiveram maiores aumentos de seus pesos do que as demais variáveis. Mantendo-se fixados os fatores proporção e erro, o fator correlação a nível alto leva a resultados até 25% melhores para o valor do SI médio.

A Tabela 3.3 mostra os percentuais de melhora do SI médio obtidos com o nível alto para o fator correlação quando comparado ao pior resultado obtido por outro nível desse mesmo fator, para dados com 100 observações; o ganho médio foi de 21%. Ganhos mais pronunciados foram obtidos nos dados com 200 (31%), 500 (39%) e 1000 observações (24%).

Tabela 3.3 – Ganho percentual para o SI médio com utilização de correlação a nível alto – 100 observações

Níveis do fator erro	Níveis do fator proporção		
	alto	nominal	baixo
alto	24%	27%	11%
nominal	26%	27%	13%
baixo	26%	27%	11%

Uma análise quanto ao fator erro revela que o mesmo não tem influência tão evidente na variação do SI médio quanto o fator correlação. Diante desse resultado, indica-se que o método conseguiu absorver os impactos gerados nos dados com a adição do erro. Uma configuração alternativa de simulação, alterando o valor do fator de escala f da soma ponderada (Eq. (3.1)), pode ser avaliada; entretanto, dada a característica aleatória dos dados gerados, espera-se que os resultados sigam a mesma tendência dos apresentados na Tabela 3.2.

A Figura 3.2 mostra o valor do SI médio nos 3 níveis do fator proporção, com os fatores correlação e erro fixados em níveis nominais, aferido no banco com 1000 observações. Percebe-se ligeira melhora no valor máximo obtido para o SI médio, mas a variação não foi tão discrepante quanto à obervada no fator correlação.

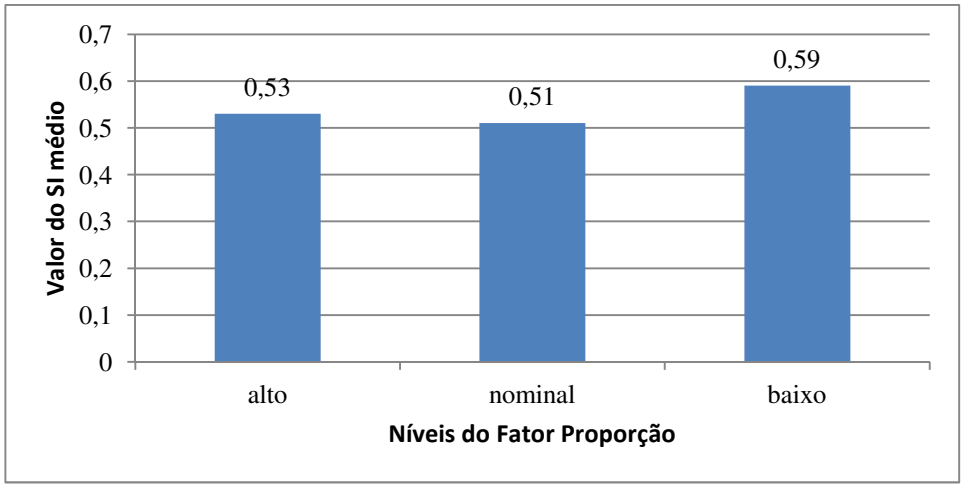

Figura 3.2 – Perfis de SI médio para o fator Proporção, a 3 níveis

Observando-se esse resultado em conjunto com a Tabela 3.2, percebe-se que o fator proporção apresenta comportamento instável, dependendo do nível do fator correlação. Quando a correlação está no nível alto, o fator proporção produziu melhores agrupamentos quando está nos níveis alto ou nominal. Quando a correlação está no nível nominal, os melhores agrupamentos são definidos com a proporção em nível baixo. Quando a correlação está no nível baixo, a proporção conduz a resultados praticamente idênticos. As melhorias constatadas são da ordem de 0,05 para os valores de SI médio, aproximadamente 10% dos valores mais baixo obtidos para o SI médio com essa configuração de níveis e fatores (0,51 com proporção nominal).

3.5 Conclusão

Ferramentas de clusterização são importantes em diversos contextos e áreas de conhecimento. A busca pela real estrutura dos dados não é uma tarefa fácil, e tem ensejado diversas abordagens. É comprovado que a geração de *clusters* representativos e distintos entre si depende de apenas um subconjunto das variáveis disponíveis e que a inserção de variáveis sem capacidade de distinção conduz a piores resultados no agrupamento de observações.

Este artigo utilizou simulação, considerando como fatores experimentais a correlação entre os dados, o nível de erro presente na obtenção desses dados e a proporção de observações utilizadas para o cálculo do IIV_p, para avaliar o método de seleção de variáveis proposto por Cervo e Anzanello (2013). A avaliação foi implementada nos seguintes passos: (1) Geração de dados simulados baseados em dados reais. (2) Aplicação da ACP sobre os dados gerados. (3) Geração de um índice de importância – IIV. (4) Definição de um intervalo de variação para o número *k* de *clusters*. (5) Realização da clusterização, com inclusão

sistemática de variáveis dirigida pelo IIV(até incluir todas as variáveis), e avaliação através do SI. (6) Incremento de k ($k = k + 1$) e retorno ao passo (5) até atingir o limite para esse número. (7) Identificação dos valores de SI, k e p (número de variáveis retidas) que conduzem à melhor clusterização. Os resultados obtidos mostram que o desempenho do método manteve os mesmos níveis obtidos pelos autores em um banco de dados real. O fator correlação é o que mais impacta o desempenho do método: a nível alto, o valor do SI médio chega a melhoras da ordem de 47% quando comparado com o pior valor obtido para outros níveis de correlação.

Em trabalhos futuros, pretende-se avaliar o impacto que a sistemática de escolha dos centroides causa no desempenho do método, assim como se vislumbra uma adaptação do algoritmo para selecionar os centroides dentre as observações disponíveis, utilizando o conceito de exemplares, disponível na literatura. Ainda, pretende-se estudar e implementar outras técnicas de clusterização indicadas para dados sabidamente não lineares.

3.6 Referências

Agard, B., Penz, B., 2009. A simulated annealing method based on a clustering approach to determine bills of materials for a large product family. **International Journal of Production Economics** 117 (2), 389-401.

Anderson, T.W., 2003. **An Introduction to Multivariate Statistical Analysis** third ed. John Wiley & Sons, Inc. Hoboken, New Jersey.

Andrews, R.L., Brusco, M.J., Currim, I.S., 2010. Amalgamation of partitions from multiple segmentation bases: A comparison of non-model-based and model-based methods. **European Journal of Operational Research** 201, 608-618.

Andrews, R.L., McNicholas, P.D., 2012. Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. **Statistics and Computing** 22 (5), 1021-1029.

Anzanello, M.J., Albin, S.L., Chaovalitwongse, W., 2009. Selecting the Best variables for classifying production batches into two quality levels. **Chemometrics and Intelligent Laboratories Systems** 97 (2), 111-117.

Anzanello, M.J., Fogliatto, F.S., 2011. Selecting the Best clustering variables for grouping mass-customized products involving workers' learning. **International Journal of Production Economics** 130 (2), 268-276.

Bellec, P., Rosa-Neto, P., Lyttelton, O.C., Benali, H., Evans, A.C., 2010. Multi-level bootstrap analysis of stable *clusters* in resting-state fMRI. **NeuroImage** 51, 1126-1139.

Brusco, M.J., 2004. Clustering binary data in the presence of masking variables. **Psychological Methods**, 9, 510-523.

Brusco, M.J., Cradit, J.D., 2001. A variable-selection heuristic for k-means clustering. **Psychometrika** 66 (2), 249-270.

Cervo, V.L., Anzanello, M.J., 2013. Sistemática de seleção de variáveis para clusterização baseada em análise de componentes principais. **Revista Gestão e Produção**. São Paulo. Aguardando publicação.

Dean, N., Raftery, A.E., 2010. Latent class analysis variable selection. **Annals of the Institute of Statistical Mathematics**, 62 (1), 11-35.

Fowlkes, E., Gnanadesikan, R., Kettenring, J., 1988. Variable selection in clustering. **Journal of Classification** 5 (2), 205-228.

Frey, B.J., Dueck, D., 2007. Clustering by passing messages between data points. **Science** 315, 972-976.

Gnanadesikan, R., Kettenring, J., Tsao, S., 1995. Weighting and selection of variables for cluster analysis. **Journal of Classification** 12 (1), 113-136.

Hair, J., Anderson, R., Tatham, R., Black, W., 1995. **Multivariate Data Analysis with Readings** fourth ed. Prentice-Hall Inc., New Jersey.

Jolliffe, I.T., 2002. **Principal Component Analysis** second ed. Springer-Verlag New York.

Kaufman, L., Rousseeuw, P., 2005. **Finding Groups in Data: an Introduction to Cluster Analysis**. Wiley Interscience, New Jersey.

Karimi, S., Hemmateenejad, B., 2013. Identification of discriminatory variables in proteomics data analysis by clustering of variables. **Analytica Chimica Acta** 767, 35-43.

Li, Y., Dong, M., Hua, J., 2008. Localized feature selection for clustering. **Pattern Recognition Letters** 29 (1), 10-18.

Liu, H.H., Ong, C.S., 2008. Variable selection in clustering for marketing segmentation using genetic algorithms. **Expert Systems with Applications** 34 (1), 502-510.

Maugis, C., Celeux, G., Martin-Magniette, M., 2009. Variable selection for clustering with Gaussian mixture models. **Biometrics** 65 (3), 701-709.

Milligan, G., 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. **Psychometrika** 45, 325-342.

Milligan, G.,1989. A validation study of a variable-weighting algorithm for cluster analysis. **Journal of Classification** 6 (1), 53-71.

Milligan, G., Cooper, M., 1988. A study of standardization of variables in cluster analysis. **Journal of Classification** 5, 181-204.

Poon, L.K.M., Zhang, N.L., Liu, T., Liu, A.H., 2013. Model based clustering of high-dimensional data: Variable selection versus facet determination. **International Journal of Approximate Reasoning** 54, 196-215.

Steinley, D., Brusco, M.J., 2008a. Selection of variables in cluster analysis: an empirical comparison of eight procedures. **Psychometrika** 73(1), 125-144.

Steinley, D., Brusco, M.J., 2008b. A new variable weighting and selection procedures for K-means cluster analysis. **Multivariate Behavioral Research** 43 (1), 77-108.

Taboada, H., Coit, D., 2008. Multi-objective scheduling problems: determination of pruned Pareto sets. **IEEE Transactions 40** (5), 552-564.

4 Terceiro Artigo: Seleção de variáveis de clusterização para o agrupamento de famílias de bateladas de produção através de remapeamento *kernel*

Victor Leonardo Cervo

Michel Jose Anzanello

Artigo a ser enviado para publicação na revista Produção

Resumo

Técnicas de clusterização buscam a formação de grupos cujas observações são homogêneas dentro de um mesmo grupo e significativamente distintas das observações inseridas em outros grupos. Em processos industriais onde a produção é caracterizada por bateladas, a definição de famílias (grupos) de bateladas com perfis semelhantes pode ajudar a definir estratégias de controle e monitoramento desses processos. Este artigo propõe uma sistemática para formação de famílias de bateladas com características similares através da utilização de análise de clusterização. Para tanto, é utilizado um artifício de remapeamento dos dados a fim de inserir nestes dados relações notadamente não lineares com vistas ao aprimoramento dos grupos formados. O mapeamento realizado por *kernels*, juntamente com uma sistemática de seleção de variáveis baseada em Análise de Componentes Principais, gerou agrupamentos 150% mais precisos quando avaliados através do *Silhouette Index* (SI), em média, frente à utilização das variáveis originais; utilizou-se, para tanto, 6% das variáveis inicialmente disponíveis, em média.

Palavras-chaves: análise de clusterização, seleção de variáveis, *kernel*, processos em batelada

Clustering variable selection for grouping production batches through *kernel* mapping

Abstract

Clustering techniques aim to find internally homogeneous groups and distinct among other groups. In industrial processes where production occurs in batches, setting similar profiles batches' families, i.e., groups, can help in defining strategies for controlling and monitoring these processes. In this paper, we propose an approach to group production batches into families relying on clustering analysis. An artifice for mapping data is used for

inserting notably non linear relations in the clustering algorithm. The mapping carried out by *kernel*s, along with a variable selection approach based on Principal Component Analysis, increased clustering precision in average 150% assessed by the Silhouette Index (SI) and used only average 6% of the original variables.

Keywords: clustering analysis, variable selection, *kernel*, batch processes

4.1 Introdução

A separação de observações em grupos distintos é objeto de estudos em diferentes áreas, justificando o elevado número de estudos com esse propósito. Dentre estes, Anzanello *et al*. (2009) utilizam a regressão por quadrados parciais mínimos (PLS - *Partial Least Squares*) para realizar a classificação de bateladas de produção em 2 grupos, em função de uma variável de resposta (nível de qualidade esperado). Steinley e Brusco (2008a) comparam oito procedimentos distintos para clusterização baseados em seleção de variáveis, recomendando os métodos de acordo com o cenário de aplicação.

Em processos da indústria química, é comum verificar-se centenas e até milhares de variáveis sendo monitoradas e utilizadas no controle desses processos. O avanço tecnológico e requisitos de competitividade permitem, e em certas ocasiões impõem, a geração de imensas bases de dados a fim de garantir processos de alto rendimento. A fim de lidar com a enorme quantidade de dados disponível, pesquisadores têm desenvolvido ferramentas de análise que demandam mínima intervenção humana, também chamadas de técnicas de aprendizado não supervisionado (Duda *et al*., 2001), das quais fazem parte as técnicas de clusterização. Além disso, estudos mostram que, de toda essa quantidade de dados disponíveis, apenas uma reduzida parte conduz à obtenção de *clusters* (grupos) relevantes (MILLIGAN, 1989; BRUSCO; CRADIT, 2001).

Com vistas à identificação dessa estrutura de *clusters*, distintas metodologias foram propostas em diversas áreas de aplicação. A técnica batizada de propagação de afinidades (Frey e Dueck, 2007) utiliza troca de 2 tipos de mensagens (disponibilidade e responsabilidade) entre as observações para definir os exemplares (centros de *clusters*, escolhidos dentre as observações); ao mesmo tempo, o número e a posição desses exemplares "naturalmente" define a quantidade e a estrutura dos *clusters*. Von Luxburg (2007) traz um guia prático sobre clusterização espectral, evidenciando os detalhes dessa técnica, enquanto

sua consistência é avaliada e discutida por von Luxburg *et al*. (2008). O presente trabalho, por sua vez, foca em métodos de particionamento dos dados, mais especificamente através do algoritmo de clusterização não hierárquico *k-means*.

A metodologia proposta neste trabalho sugere a utilização de funções *kernel* como ferramenta auxiliar para a clusterização. O objetivo é realizar um remapeamento dos dados para outro espaço, onde a nova representação dos dados possibilite a formação de *clusters* mais consistentes. Apesar de utilizar *kernel* para remapeamento e *k-means* como algoritmo de clusterização, a proposta deste trabalho difere do tradicional *kernel k-means* proposto pela literatura: aqui o mapeamento será integrado ao método proposto por Cervo e Anzanello (2013a), realizado antes da aplicação da Análise de Componentes Principais, como forma de inserir na análise relações não lineares. Em sua proposição original, Dhillon *et al*. (2004) utilizam um *kernel* para computar as distâncias entre as observações, deixando de utilizar a distância euclidiana, implementando o tradicional *kernel k-means* reportado pela literatura.

O presente artigo está organizado como segue, além desta introdução. A seção 4.2 faz uma breve revisão da fundamentação teórica sobre seleção de variáveis para clusterização e funções de *kernel*. A Seção 4.3 apresenta o método proposto, enquanto a Seção 4.4 reporta os resultados numéricos da sistemática proposta aplicada a bancos reais, bem como a discussão desses resultados. Por fim, as conclusões e os direcionamentos futuros são apresentados na Seção 4.5.

4.2 Fundamentação teórica

4.2.1 Seleção de variáveis para clusterização

A formação de grupos compostos por observações similares entre si e diferentes das alocadas a grupos vizinhos é o objetivo das técnicas de clusterização (Kaufman e Rousseeuw, 2005). Existem diferentes classes de algoritmos utilizados com esse propósito, sendo a divisão usual feita em (1) algoritmos hierárquicos e (2) não hierárquicos (Hair *et al*., 1995). Os métodos hierárquicos definem hierarquias entre as observações, representando essas relações através de um dendograma, estrutura semelhante a uma árvore. Esse tipo de método é bastante utilizado para a determinação do número de *clusters* em que as observações devem ser alocadas, através de um corte nos ramos do dendograma. Os métodos não hierárquicos não definem hierarquias; eles são chamados métodos de particionamento, pois dividem as observações em um número definido de grupos. O mais conhecido algoritmo desse tipo, o *k-means*, é um dos mais utilizados por conta de sua velocidade de convergência. Ele realiza a

inserção das observações em *k clusters*, através da escolha aleatória de centros de grupos (centroides) na primeira iteração, e da minimização de uma função objetivo apoiada na soma global das distâncias de cada observação ao centroide de seu *cluster*. Os centroides são recalculados a cada iteração. Contudo, o *k-means* é altamente dependente da escolha inicial desses centroides, garantindo apenas a obtenção de ótimos locais. Steinley (2006) apresenta uma investigação sobre essa limitação, avaliando diversos fatores que conduzem ao ótimo local.

A qualidade do procedimento de clusterização pode ser calculada através do *Silhouette Index* (SI), que informa quanto uma observação é semelhante a outras alocadas no mesmo *cluster* frente a outras alocadas no *cluster* vizinho (Anzanello e Fogliatto, 2011). De tal forma, é possível avaliar o desempenho geral da clusterização através da média dos SI's das observações agrupadas (TABOADA; COIT, 2008; CERVO; ANZANELLO, 2013a; 2013b).

Diversos estudos sugerem que agrupamentos mais representativos são obtidos quando um conjunto reduzido de variáveis é utilizado, visto que a inserção de variáveis ruidosas pode degradar significativamente a qualidade do procedimento (Milligan, 1980; Li *et al*., 2008; Maugis *et al*., 2009). Brusco (2004) recomenda a anulação do efeito dessas variáveis ruidosas através da seleção de variáveis. Alguns autores utilizam ferramentas de análise multivariada para auxiliar na tarefa de selecionar variáveis: Fowlkes *et al*. (1988) e Gnanadesikan *et al* (1995) utilizam Análise de Variância Multivariada; Anzanello *et al*. (2009) utilizam regressão por mínimos quadrados parciais (*Partial Least Squares* – PLS); Cervo e Anzanello (2013a, 2013b) utilizam Análise de Componentes Principais (ACP). Por sua vez, Steinley e Brusco (2008b) consideram informações relativas à variância como base para a seleção de variáveis.

Seguindo a linha de seleção de variáveis, Raftery e Dean (2006) consideram que as variáveis formam subgrupos, relevantes e irrelevantes, e a cada iteração do algoritmo é avaliada a possibilidade de trocar variáveis de grupos. Dean e Raftery (2010) sistematizam a seleção de variáveis com dados categóricos, aplicando o método em mapeamento genético. Os dois estudos se aplicam ao contexto de modelos de misturas finitas, considerando que a população contém subgrupos com distribuições e proporções próprias. Ainda, Bessaoud *et al*. (2012) adaptam a ideia de seleção de variáveis fazendo a clusterização de variáveis com o objetivo de identificar padrões de dieta relacionados ao risco de desenvolver câncer de mama.

4.2.2 Funções *kernel*

As funções de *kernel* são, em sua definição mais básica, medidas de similaridade entre observações. Considerando duas observações x e x' pertences a um conjunto X, define-se uma medida de similaridade entre os elementos de X na forma da Eq. (4.1) (SCHÖLKOPF; SMOLA, 2002):

$$k : X \times X \rightarrow \mathbf{IR}$$

$$(x, x') \rightarrow k(x, x'), \tag{4.1}$$

ou seja, uma função que recebe como parâmetros de entrada 2 observações e retorna um número real que caracteriza a similaridade entre essas observações. Assume-se, de maneira geral, que a função que mede similaridades é simétrica, com $k(x, x') = k(x', x)$.

Uma medida simples e bastante utilizada como medida de similaridade é o produto escalar canônico, conforme a Eq. (4.2) (SCHÖLKOPF; SMOLA, 2002),

$$\langle \mathbf{x}, \mathbf{x}' \rangle := \sum_{i=1}^{N} [\mathbf{x}]_i [\mathbf{x}']_i \tag{4.2}$$

onde $\mathbf{x}$ e $\mathbf{x}'$ são vetores tais que $\mathbf{x}, \mathbf{x}' \in \mathbf{IR}^{\mathbf{N}}$, $[\mathbf{x}]_i$ representa a i-ésima entrada do vetor $\mathbf{x}$. Geometricamente, o produto escalar canônico, também chamado de produto interno, calcula o cosseno do ângulo entre os vetores $\mathbf{x}$ e $\mathbf{x}'$, desde que os mesmos estejam normalizados com comprimento 1. A norma de um vetor pode ser calculada fazendo-se

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}. \tag{4.3}$$

A distância entre 2 vetores pode ser calculada como o comprimento do vetor diferença. Resumidamente, ser capaz de calcular produtos internos é poder utilizar todas as construções geométricas que utilizam em sua formulação ângulos, comprimentos e distâncias. Nesse contexto inserem-se as funções de mapeamento.

É possível que os dados de entrada não estejam representados em um espaço em que exista produto interno. Para se utilizar o produto interno como medida de similaridade, faz-se necessário um mapeamento das observações como vetores em um espaço K, o qual não precisa ser $\mathbf{IR}^{\mathbf{N}}$. Para isso, utiliza-se um mapa como na Eq. (4.4)

$$\Phi : X \rightarrow K$$

$$x \rightarrow \mathbf{x} := \Phi(x). \tag{4.4}$$

A aplicação de mapas permite a utilização de formas genéricas de medidas de similaridades; com esse objetivo, é usual que Φ seja um mapa não linear (SCHÖLKOPF; SMOLA, 2002).

Schölkopf e Smola (2002) apontam 3 vantagens de realizar o mapeamento de observações para *K* através de Φ: (i) É possível definir o produto interno em *K* como uma medida de similaridade, ou seja,

$$k(x, x') = \langle \mathbf{x}, \mathbf{x}' \rangle = \langle \Phi(x), \Phi(x') \rangle. \quad (4.5)$$

(ii) É possível trabalhar geometricamente com as observações, o que implica a possibilidade de desenvolver algoritmos usando álgebra linear e geometria analítica, e (iii) o mapeamento não está restrito ao caso em que o espaço *X* não tenha produtos internos; pode-se fazer o mapeamento para qualquer espaço de interesse, principalmente utilizando um mapa não linear, a fim de representar os dados da maneira que seja mais interessante para um problema específico. De tal forma, pode-se escolher um mapa Φ não linear capaz de mapear os dados *x,* do espaço *X* dos dados, para a representação **x**, do espaço *K* dos atributos, de tal forma que problemas não linearmente resolvíveis em *X* passem a ser em *K*, através da separação por um hiperplano.

Huang *et al*. (2006) afirmam que o cálculo dos produtos internos $\langle \Phi(x), \Phi(x') \rangle$ em *K* pode ter alto custo computacional, tendo em vista que esse espaço pode ter uma dimensionalidade muito maior do que *X*. Entretanto, como tem-se $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$ de acordo com a Eq. 4.5 e a função $k(x, x')$ está no espaço *X* das observações, o cálculo dos produtos internos pode ser evitado, utilizando-se a função de *kernel* $k(x, x')$, no espaço de observações.

A escolha apropriada da função de *kernel* constitui-se num ponto importante de análise, sendo o desenvolvimento de novos *kernels* um tópico de pesquisa recente. Abe (2010) e Schölkopf e Smola (2002) destacam as funções de *kernel* mais usuais:

- Lineares, com $k(x, x') = x^{\mathrm{T}} x'$ (4.6)
- Polinomiais, de grau *d*, com $k(x, x') = (x^{\mathrm{T}} x' + 1)^d$ (4.7)
- Gaussianos, com $k(x, x') = e^{-\frac{\|x - x'\|^2}{2\sigma^2}}$, com $\sigma \neq 0$ (4.8)
- Sigmóides, com $k(x, x') = \tanh(\kappa \langle x, x' \rangle + \vartheta)$, (4.9)

  onde $\kappa > 0$ e $\vartheta < 0$.

Estudos têm recentemente avaliado a utilização de *kernels* e mapeamento para espaços de atributos. Filippone *et al*. (2008) investigam a adaptação de métodos de particionamento, como o *k-means*, através da utilização de *kernels*; os resultados obtidos mostram que a aplicabilidade dos métodos ainda é problemática, tendo em vista o alto custo computacional. Domenicone *et al*. (2011) estudam técnicas de otimização para os parâmetros da função *kernel*, em um contexto de clusterização semi-supervisionada, guiada através de restrições dos tipos deve-ligar e não-pode-ligar. Por sua vez, Baghshah e Shouraki (2011) desenvolveram um método de aprendizado com métrica não linear que obtém as funções de *kernel* a partir das restrições para clusterização e da topologia dos dados; os resultados sugerem que o método proposto tem potencial de uso, resolvendo o problema de otimização de forma mais eficiente que outros métodos existentes.

4.3 Método

A sistemática proposta utiliza o mapeamento por funções de *kernel* como etapa de pré-processamento dos dados (representando-os em outro espaço), para então aplicar o método proposto por Cervo e Anzanello (2013a) a fim de obter *clusters* representativos de famílias de bateladas de produção através de seleção de variáveis. A metodologia é implementada em 3 passos: (1) Realizar o pré-processamento dos dados utilizando as funções *kernel*. (2) Proceder à normalização dos dados. (3) Realizar a seleção de variáveis e clusterização de acordo com Cervo e Anzanello (2013a). Esses passos são detalhados na sequência.

4.3.1 Passo 1 – Pré-processamento dos dados através de funções *kernel*

Inicialmente, os dados devem ser transformados através de uma função *kernel*. A transformação tem por objetivo inserir no contexto da análise eventuais relações não lineares, utilizando para isso um *kernel* polinomial. Espera-se que a nova representação dos dados promova a formação de grupos mais precisos quando comparados à sistemática proposta por Cervo e Anzanello (2013a).

Para efeitos de avaliação, este trabalho utiliza 2 *kernels*: um para a transformação $\mathbf{X^3}$, outro para a transformação $\mathbf{X^{1/3}}$. A transformação com expoente fracionário com denominador ímpar (neste caso, equivalente à raiz cúbica) teve a escolha motivada no fato de que, para algumas variáveis, os valores dos dados podem ter sinal negativo. Caso fosse escolhida uma potência fracionária com denominador par, ½ por exemplo, os novos dados poderiam se tornar números complexos, descaracterizando o banco original.

4.3.2 Passo 2 – Normalização dos dados

Os dados remapeados são, neste passo, escalonados para o intervalo [0,10] para igualar a ordem de magnitude de todas as variáveis, a fim de garantir consistência na Análise de Componentes Principais (ACP) e na clusterização (Milligan e Cooper, 1988; Steinley, 2004; Anzanello e Fogliatto, 2011). A ACP é uma ferramenta de análise multivariada que altera o espaço original de representação dos dados para um espaço de eixos ortogonais entre si (componentes). Esses componentes são os autovetores da matriz de covariância dos dados. Jolliffe (2002) e Anderson (2003) apresentam detalhadamente a derivação dos componentes. A análise fornece como parâmetros de saída os componentes (com os pesos de cada variável na sua composição) e os novos dados; os autovalores da matriz de covariância também podem ser de interesse, tendo em vista que representam a variância associada a cada componente principal (CERVO; ANZANELLO, 2013a).

4.3.3 Passo 3 – Clusterização das observações e avaliação do desempenho

Nesse passo, procede-se conforme as proposições de Cervo e Anzanello (2013a): (1) Aplica-se ACP nos dados transformados e reescalonados. (2) Gera-se um índice de importância de variável (IIV), o qual é utilizado como referência na inclusão sistemática de variáveis no subconjunto de variáveis de clusterização. (3) Define-se um intervalo de variação para o número de *clusters*. (4) Os procedimentos de clusterização iniciam com as 2 variáveis mais importantes; as demais são inseridas sistematicamente obedecendo à ordem obtida pelo IIV. A cada nova rodada, o SI de todas as observações é computado, permitindo a comparação do desempenho do procedimento para diferentes números de variáveis e *clusters* através do cálculo do SI médio. (5) Após inserir todas as variáveis, altera-se o número de *clusters* e retorna-se à etapa 4. (6) Avaliam-se os resultados obtidos e indicam-se o número *k* de *clusters* e o conjunto de variáveis selecionadas que obtêm o maior valor para o SI médio, denotando melhores agrupamentos.

Os resultados globais obtidos com a utilização de funções de *kernel* são, então, comparados aos resultados obtidos com a utilização dos dados em seu formato original, a fim de avaliar o desempenho do remapeamento e seu potencial de utilização.

4.4 Exemplos numéricos

A sistemática proposta foi aplicada a 3 bancos de dados da indústria química, conforme descritos na Tabela 4.1. O banco ADPN se refere à produção de um subcomponente no processo de produção do *nylon*. O banco LATEX foi coletado em um estágio de

polimerização da fabricação de látex. O banco SPIRA foi obtido em um processo da indústria farmacêutica para produção de um antibiótico. Maiores detalhes acerca de tais bancos podem ser obtidos em Gauchi e Chagnon (2001).

Tabela 4.1 – Descrição dos Bancos de Dados utilizados

Banco de Dados	Número de variáveis	Número de observações
ADPN	100	71
LATEX	117	262
SPIRA	96	145

A implementação da metodologia proposta foi realizada através do aplicativo MATLAB®, versão 7.0.0.19920 (R14). A função criada para realizar as análises apresenta como parâmetros de saída os valores do SI médio para os dados no espaço original e nos espaços remapeados, para um banco de dados. Os tempos de execução consideram a leitura dos bancos de dados originais e todo o processamento necessário para obter os valores de SI médios para cada conjunto possível de números de *clusters* gerados e variáveis selecionadas para clusterização: para o banco ADPN, o tempo total foi de 22 segundos aproximadamente; para o banco LATEX, foi de 3 minutos e 15 segundos; para o banco SPIRA, o tempo total foi de, aproximadamente, 55 segundos.

A Figura 4.1 apresenta o perfil de SI médio para o Banco ADNP à medida que as variáveis de clusterização são inseridas no procedimento para 2 *clusters* e *kernels* distintos. Percebe-se que a inserção de variáveis degrada o valor do valor médio do índice SI, confirmando que um subconjunto reduzido das variáveis originais conduz a melhores agrupamentos de bateladas produtivas. Percebe-se uma alternância no desempenho de clusterização entre os dois *kernel*s testados que, contudo, não supera o desempenho da sistemática proposta por Cervo e Anzanello (2013a) utilizada sobre os dados sem remapeamento.

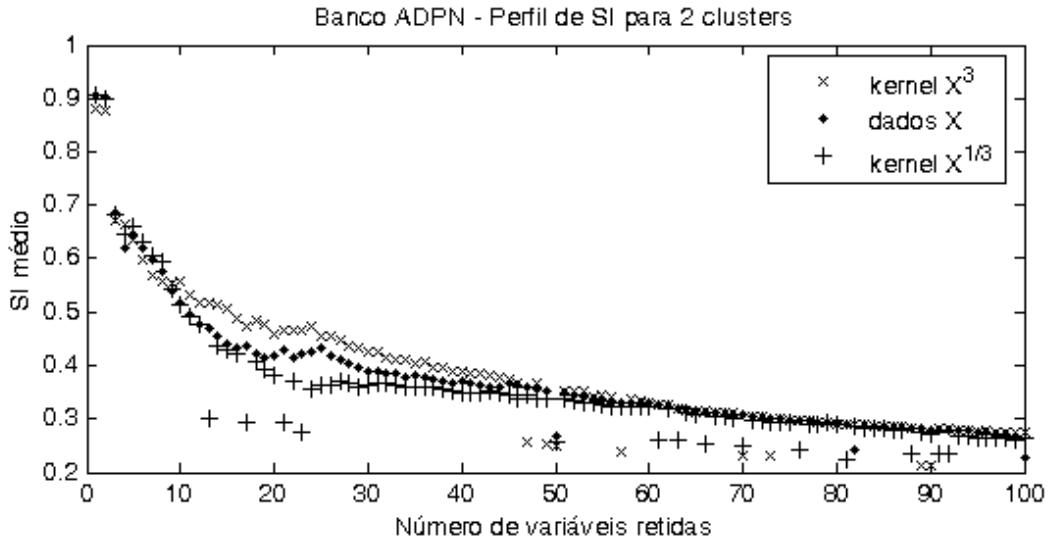

Figura 4.1 – Perfis de SI médio para o Banco ADPN na formação de 2 *clusters*

A Figura 4.2 refere-se ao banco LATEX. Pode-se observar que o método proposto por Cervo e Anzanello (2013a), sem a utilização de remapeamento, é capaz de identificar a formação preferencial de 2 *clusters* com base em um limitado número de variáveis, tendo em vista os elevados valores de SI médio obtidos. Com apenas 3 variáveis retidas, o desempenho global da clusterização em 2 *clusters* gerou SI médio igual a 0,94, refletindo grande precisão na alocação das observações aos grupos. Para a divisão em 3 grupos, os resultados não se mostraram tão adequados quanto os da divisão em 2 grupos, tendo em vista a própria estruturação do banco, mas ainda assim os valores médios de SI obtidos superam 0,7 para poucas variáveis.

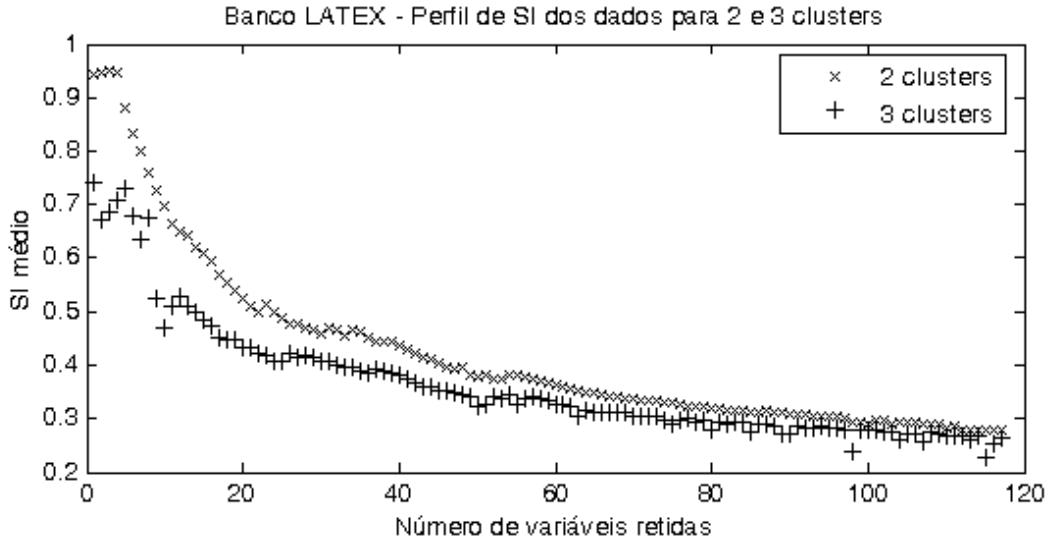

Figura 4.2 – Perfis de SI médio para o Banco LATEX na formação de 2 e 3 *clusters*

A Tabela 4.2 apresenta os valores máximos de SI médio quando realizados os remapeamentos $\mathbf{X}^{1/3}$, $\mathbf{X}^{3}$ e os dados **X**, para 2, 3, 4 e 5 *clusters*. Os valores em negrito ressaltam o melhor desempenho obtido para cada banco em cada número de *clusters*. Nos casos em que há valores iguais na Tabela 4.2, o desempate foi com base nas demais casas decimais, não apresentadas por restrição de espaço na tabela. Observa-se que os dados originais apresentaram melhor qualidade na clusterização em apenas 3 oportunidades, enquanto o remapeamento para $\mathbf{X}^{1/3}$ foi melhor em 4 cenários e o remapeamento para $\mathbf{X}^{3}$ foi melhor em 5 cenários.

Tabela 4.2 – Valores máximos para o SI médio para distintos *kernels* e número de clusters

Bancos de dados	$k = 2$			$k = 3$			$k = 4$			$k = 5$		
	$\mathbf{X}^{1/3}$	**X**	$\mathbf{X}^{3}$	$\mathbf{X}^{1/3}$	**X**	$\mathbf{X}^{3}$	$\mathbf{X}^{1/3}$	**X**	$\mathbf{X}^{3}$	$\mathbf{X}^{1/3}$	**X**	$\mathbf{X}^{3}$
ADPN	0,90	**0,90**	0,87	0,79	0,80	**0,86**	0,79	**0,80**	0,80	**0,73**	0,73	0,73
LATEX	0,86	**0,94**	0,89	0,87	0,72	**0,89**	**0,92**	0,68	0,81	**0,81**	0,66	0,77
SPIRA	0,59	0,55	**0,85**	0,66	0,55	**0,80**	**0,89**	0,50	0,85	0,77	0,51	**0,80**

Percebe-se a grande recuperação de informações sobre a estrutura dos bancos ADPN e LATEX para 2 *clusters*: tanto com a utilização das funções de *kernel*, quanto dos dados originais, a sistemática de seleção de variáveis obteve bons resultados. A utilização dos *kernels* para números de *clusters* maiores do que 2 obteve agrupamentos substancialmente melhores do que a utilização dos dados originais para o banco LATEX e SPIRA; especificamente para 4 *clusters*, o mapeamento realizado pelo *kernel* $\mathbf{X}^{1/3}$ gerou SI's médios maiores do que 0,85, bastante superiores aos obtidos com a utilização dos dados originais. Para o banco ADPN, a utilização dos *kernels* obteve um resultado ligeiramente melhor para 3 *clusters* e resultados similares para 4 e 5 *clusters*.

A Tabela 4.3 apresenta, nos mesmos moldes da Tabela 4.2, o ganho percentual do maior valor de SI médio obtido pela utilização do subconjunto de variáveis de clusterização recomendado pela sistemática. Observa-se que o banco SPIRA é o que apresenta maiores divergências nos valores de ganho aferidos, sendo responsável pelos maiores e menores níveis percentuais de ganho constatados. Verifica-se que a função *kernel* $\mathbf{X}^{1/3}$ apresentou significativos percentuais de melhora dentro da sistemática de seleção de variáveis; contudo, observando-se em conjunto a Tabela 4.2, a utilização desse *kernel* levou a melhores resultados do que o *kernel* $\mathbf{X}^{3}$ em 5 de 12 possibilidades, e em 4 dessas 5 conduziu o SI médio ao melhor resultado obtido na análise.

Tabela 4.3 – Ganho percentual dos SI's médios obtidos pela sistemática de seleção de variáveis

Bancos de dados	$k = 2$			$k = 3$			$k = 4$			$k = 5$		
	$\mathbf{X}^{1/3}$	$\mathbf{X}$	$\mathbf{X}^3$	$\mathbf{X}^{1/3}$	$\mathbf{X}$	$\mathbf{X}^3$	$\mathbf{X}^{1/3}$	$\mathbf{X}$	$\mathbf{X}^3$	$\mathbf{X}^{1/3}$	$\mathbf{X}$	$\mathbf{X}^3$
ADPN	246	309	222	182	175	230	203	166	280	329	265	204
LATEX	230	248	256	262	176	25	283	257	224	285	153	83
SPIRA	195	175	80	371	266	73	709	257	84	600	292	81
Média	223,7	244,0	186,0	271,7	205,7	109,3	398,3	226,7	196,0	404,7	236,7	122,7

A Tabela 4.4 apresenta os percentuais de variáveis retidas pela sistemática, os quais estão atrelados aos valores de SI médio apresentados na Tabela 4.2.

Tabela 4.4 – Percentual de variáveis retidas pela sistemática de seleção de variáveis

Bancos de dados	$k = 2$			$k = 3$			$k = 4$			$k = 5$		
	$\mathbf{X}^{1/3}$	$\mathbf{X}$	$\mathbf{X}^3$	$\mathbf{X}^{1/3}$	$\mathbf{X}$	$\mathbf{X}^3$	$\mathbf{X}^{1/3}$	$\mathbf{X}$	$\mathbf{X}^3$	$\mathbf{X}^{1/3}$	$\mathbf{X}$	$\mathbf{X}^3$
ADPN	2,00	2,00	2,00	2,00	2,00	2,00	3,00	3,00	2,00	2,00	3,00	4,00
LATEX	4,27	2,56	3,42	2,56	4,27	3,42	1,71	2,56	3,42	2,56	3,42	5,98
SPIRA	2,08	2,08	93,75	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08
Média	2,78	2,21	33,06	2,21	2,78	2,50	2,26	2,55	2,50	2,21	2,83	4,02

Observa-se que os percentuais de variáveis retidas ficam abaixo de 5% na maioria dos casos, com uma média geral de 5,16% devida a uma exceção: o mapeamento para $\mathbf{X}^3$ no banco SPIRA, na formação de 2 *clusters*. Praticamente todas as variáveis foram retidas para obter um SI de 0,85, conforme a Tabela 4.4.

4.5 Conclusões

A obtenção de grupos de observações distintos entre si, mas com afinidade interna entre seus integrantes, é uma tarefa de grande interesse em várias áreas de conhecimento. No contexto da indústria química, mais do que a simples classificação de bateladas de produção, o grande desafio é estabelecer relações de níveis de qualidade entre essas bateladas.

Neste artigo foi proposta uma sistemática que, antes de realizar os procedimentos de clusterização com base nos parâmetros da ACP, faz o remapeamento das observações, utilizando *kernels*, a fim de gerar uma nova representação do conjunto de dados. A inserção de relações não lineares teve como objetivo obter melhores agrupamentos para bateladas de produção.

O método proposto foi aplicado a 3 bancos distintos, apresentando bons resultados. Comparado à utilização da sistemática sem remapeamento, a utilização dos *kernels* elevou, na maior parte dos casos, a qualidade dos agrupamentos, de onde conclui-se que a sistemática com remapeamento tem grande potencial de uso. Dos 3 bancos testados em 4 cenários

(distinto número de *clusters*), a utilização dos *kernels* mostrou-se superior em 9 das 12 possibilidades.

Ressalta-se que o objetivo da sistemática proposta é agrupar as observações de forma a maximizar o SI. Estudos futuros podem contemplar o desenvolvimento de uma sistemática que compatibilize elevados valores de SI com reduzido percentual de variáveis retidas. Para tanto, sugere-se a adoção de um critério de distância mínima a um ponto ótimo hipoteticamente arbitrado (por exemplo, SI médio igual a 1 e percentual de variáveis retidas igual a 0,01). Tal proposição priorizaria a seleção de subconjuntos que originam SI's elevados e com poucas variáveis retidas.

4.6 Referências

Abe, S., 2010. **Support Vector Machines for Pattern Recognition** second ed. Springer-Verlag, London.

Anderson, T.W., 2003. **An Introduction to Multivariate Statistical Analysis** third ed. John Wiley & Sons, Inc. Hoboken, New Jersey.

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

Anzanello, M.J., Fogliatto, F.S., 2011. Selecting the best clustering variables for grouping mass-customized products involving workers' learning. **International Journal of Production Economics** 130 (2), 268-276.

Baghshah, M.S., Shouraki, S.B., 2011. Learning low-rank *kernel* matrices for constrained clustering. **Neurocomputing** 74, 2201-2211.

Bessaoud, F., Tretarre, B., Daurès, J.P., Gerber, M., 2012. Identification of dietary patterns using two statistical approaches and their association with breast cancer risk: a case-control study in southern France. **Annals of Epidemiology** 22 (7), 499-510.

Brusco, M.J., 2004. Clustering binary data in the presence of masking variables. **Psychological Methods**, 9, 510-523.

Brusco, M.J., Cradit, J.D., 2001. A variable-selection heuristic for k-means clustering. **Psychometrika** 66 (2), 249-270.

Cervo, V.L., Anzanello, M.J., 2013a. Sistemática de seleção de variáveis para clusterização baseada em análise de componentes principais. **Revista Gestão e Produção**. São Paulo. Aguardando publicação.

Cervo, V.L., Anzanello, M.J., 2013b. Avaliação de robustez de uma sistemática de seleção de variáveis para clusterização através de experimentos de simulação. **Revista Gestão e Produção**. São Paulo. Aguardando publicação.

Dean, N., Raftery, A.E., 2010. Latent class analysis variable selection. **Annals of the Institute of Statistical Mathematics**, 62 (1), 11-35.

Dhillon, I.S., Guan, Y., Kulis, B., 2004. *Kernel* k-means, spectral clustering and normalized cuts. **Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, 551-556.

Domenicone, C., Peng, J., Yan, B., 2011. Composite *kernel*s for semi-supervised clustering. **Knowledge and Information Systems** 28 (1), 99-116.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. **Pattern Classification** second ed, Wiley-Interscience, New York.

Filippone, M., Camastra, F., Masulli, F., Rovetta, S., 2008. A survey of *kernel* and spectral methods for clustering. **Pattern Recognition** 41 (1), 176-190.

Fowlkes, E., Gnanadesikan, R., Kettenring, J., 1988. Variable selection in clustering. **Journal of Classification** 5 (2), 205-228.

Frey, B.J., Dueck, D., 2007. Clustering by passing messages between data points. **Science** 315, 972-976.

Gnanadesikan, R., Kettenring, J., Tsao, S., 1995. Weighting and selection of variables for cluster analysis. **Journal of Classification** 12 (1), 113-136.

Gauchi, J.P., Chagnon, P., 2001. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. **Chemometrics Intelligent Laboratory Systems** 58, 171-193.

Hair, J., Anderson, R., Tatham, R., Black, W., 1995. **Multivariate Data Analysis with Readings** fourth ed. Prentice-Hall Inc., New Jersey.

Huang, T., Kecman, V., Kopriva, I., 2006. ***Kernel* based algorithms for mining huge data sets, Supervised, Semi-supervised, and Unsupervised learning**. Springer-Verlag, Berlin, Heidelberg.

Kaufman, L., Rousseeuw, P., 2005. **Finding Groups in Data: an Introduction to Cluster Analysis**. Wiley Interscience, New Jersey.

Li, Y., Dong, M., Hua, J., 2008. Localized feature selection for clustering. **Pattern Recognition Letters** 29 (1), 10-18.

Maugis, C., Celeux, G., Martin-Magniette, M., 2009. Variable selection for clustering with Gaussian mixture models. **Biometrics** 65 (3), 701-709.

Milligan, G., 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. **Psychometrika** 45, 325-342.

Milligan, G.,1989. A validation study of a variable-weighting algorithm for cluster analysis. **Journal of Classification** 6 (1), 53-71.

Milligan, G., Cooper, M., 1988. A study of standardization of variables in cluster analysis. **Journal of Classification** 5, 181-204.

Raftery, A.E., Dean, N., 2006. Variable selection for model-based clustering. **Journal of the American Statistical Association** 101, 168-178.

Schölkopf, B., Smola, A.J., 2002. **Learning with *Kernel*s: Support Vector Machines, Regularization, Optimization, and Beyound**. The MIT Press, Cambridge, Massachussets.

Steinley, D., 2004. Standardizing variables in *K*-means clustering. In D. Banks, L. House, F.R. McMorris, P. Arabie, & W. Gaul (Eds.), **Classification, clustering, and data mining applications** (pp. 53-60). New York: Springer.

Steinley, D., 2006. Profiling local optima in *K*-means clustering: developing a diagnostic technique. **Psychological Methods** 11 (2), 178-192.

Steinley, D., Brusco, M.J., 2008a. Selection of variables in cluster analysis: an empirical comparison of eight procedures. **Psychometrika** 73 (1), 125-144.

Steinley, D., Brusco, M.J., 2008b. A new variable weighting and selection procedure for K-means cluster analysis. **Multivariate Behavioral Research 43** (1), 77-108.

Taboada, H., Coit, D., 2008. Multi-objective scheduling problems: determination of pruned Pareto sets. **IEEE Transactions 40** (5), 552-564.

von Luxburg, U., 2007. A tutorial on spectral clustering. **Statistics and Computing**, 17 (4), 395-416.

von Luxburg, U., Belkin, M., Bousquet, O., 2008. Consistency of spectral clustering. **The Annals of Statistics** 36 (2), 555-586.

5 Considerações finais

Este capítulo apresenta as conclusões do trabalho, além de apontar sugestões para trabalhos futuros.

5.1 Conclusões

O trabalho apresentado teve como principal objetivo propor sistemáticas de seleção de variáveis para clusterização.

A revisão bibliográfica realizada nos três artigos permitiu que o primeiro objetivo específico fosse alcançado; foram apresentadas as fundamentações teóricas de análise de clusterização com foco em seleção de variáveis.

O primeiro artigo atingiu dois objetivos específicos, uma vez que apresentou um novo índice de importância das variáveis a partir dos parâmetros da Análise de Componentes Principais (ACP), além de ter integrado tal ferramenta a técnicas de clusterização.

O segundo artigo atingiu outro objetivo específico ao avaliar a sistemática proposta no primeiro artigo através de experimentos de simulação em que foi avaliada a qualidade da clusterização através da variação dos níveis dos fatores considerados: correlação, ruído e proporção dos dados.

O terceiro artigo alcança o último objetivo específico ao apresentar uma modificação na sistemática proposta no primeiro artigo, incluindo o remapeamento dos dados através de funções de *kernel* como etapa de pré-análise.

Portanto, acredita-se que todos os objetivos específicos foram alcançados e, portanto, pode-se dizer que o objetivo principal deste trabalho foi igualmente alcançado.

O primeiro artigo apresentou a fundamentação teórica sobre análise de clusterização, dando especial atenção a métodos de seleção de variáveis, e ACP. Foi proposta uma nova sistemática de seleção de variáveis, através da definição de um novo índice de importância de variáveis para clusterização, baseado em parâmetros oriundos da ACP. A sistemática é do tipo *forward* e inclui sistematicamente as variáveis no conjunto de variáveis de clusterização seguindo a ordem definida pelo índice. Os resultados obtidos, avaliados através do *Silhouette Index*, mostraram a retenção média de 9% das variáveis, com uma melhora média de 100% para o valor do SI quando comparado com a utilização do total de variáveis disponíveis. Conclui-se que o método, apesar de apoiar-se em ferramentas simples, conduz a melhorias

significativas no potencial de clusterização de observações, além de reter um percentual significativamente reduzido de variáveis.

O segundo artigo utilizou simulação para avaliar a robustez do método proposto no primeiro artigo. O desempenho da clusterização é avaliado através da análise de fatores como correlação e ruído nos dados, além da proporção de observações disponíveis. Os resultados indicaram que a sistemática de seleção de variáveis proposta tem bom desempenho em diferentes cenários, mostrando-se adequada para ser utilizada como ferramenta auxiliar para a análise de clusterização.

O terceiro artigo apresentou uma modificação na metodologia proposta no primeiro artigo ao inserir na análise dos dados uma etapa de mapeamento através de funções de *kernel*, com a intenção de incluir relações não lineares entre os dados. Tanto a metodologia originalmente proposta quanto a utilização de *kernels* conduziram a resultados satisfatórios. Com incrementos no valor do *Silhouette Index* da ordem de 150%, o remapeamento por *kernels* utilizou em torno de 6% das variáveis disponíveis, mostrando-se uma ferramenta com potencial de utilização. De tal forma, conclui-se que a integração de funções *kernel* à sistemática originalmente proposta por Cervo e Anzanello (2013) conduz à formação mais precisa de agrupamentos de observações.

5.2 Sugestões para trabalhos futuros

Como estudos que podem expandir as proposições desta dissertação, sugerem-se as seguintes pesquisas:

- Testar a aplicabilidade do índice de importância proposto em conjunto com outras sistemáticas de seleção de variáveis existentes na literatura;
- Propor novas maneiras de definição de centroides para o algoritmo *k-means*, embasadas nas informações fornecidas pela ACP;
- Estudar a relação entre as metodologias propostas nessa dissertação e as metodologias existentes mais indicadas pela literatura para dados com relações não lineares.
- Estudar a definição de critérios adicionais para a escolha do melhor número de *clusters* e de variáveis retidas, baseados no SI.