# Estudo de Técnicas Multivariadas Para Seleção de Variáveis em Grandes Bancos de Dados: Uma Aplicação Envolvendo Dados de Inibição ($IC_{50}$)

**Jaciele de Jesus Oliveira[1], Antônio Luiz Silveira Vilanova Costa[2],João Batista filgueira costa[3], Guilherme Rocha Moreira[4] , Nivan Bezerra da Costa Júnior[5], Carlos Raphael Araújo Daniel[6]**

## 1. Introdução

Em estudos envolvendo grandes bancos de dados contendo dezenas ou até centenas de variáveis correlacionadas entre si, a análise destas se torna mais complicada, pois a dificuldade no reconhecimento de padrões aumenta, assim como a possibilidade de erros, o custo computacional e o grau decomplexidade na interpretação dos resultados. As técnicas de análise multivariada surgem como um meio eficiente na seleção de variáveis, visto que tais técnicas são utilizadas para identificar um número reduzido de variáveis relevantes que melhor descrevam o banco de dados analisado sem prejuízo de informação útil (MANLY, 2008).

No presente estudo serão abordados trêsmétodos de análise multivariada de dados: Regressão por Componentes Principais (PCR – *Principal ComponentsRegression*),Regressão por Mínimos Quadrados Parciais (PLSR – *PartialLeastSquaresRegression*) e Análise de agrupamentos (*Cluster analisys*) para descrever a associação entre diversas variáveis e o $IC_{50}$, que é uma medida da concentração necessária para que uma dada substância seja capaz de inibir 50% da atividade química ou biológica de um alvo de interesse, de modo que quantomaior o valor do $IC_{50}$, maior será a quantidade necessária para inibir um determinado processo pela metade, portanto, o ideal seria que este valor fosse o menor possível, pois assim o composto avaliado seria mais ativo.

A Análise de Componentes Principais tem por objetivo resumir os dados contidos numa tabela com *p*variáveis numéricas medidas em *n*indivíduos. Este tipo de análise é considerado um método fatorial, pois a redução do número de variáveis não se faz simplesmente excluindo algumas variáveis e mantendo outras, mas pela construção de novas variáveis sintéticas, obtidas pela combinação linear das variáveis inicias, por meio dos fatores (BOUROCHE, 1982).

Desenvolvida em meados dos anos 60 por Herman O. A. Wold, a regressão PLS foi originalmente construída para o uso no campo da econometria, mas foi adotada pelo campo da quimiometria. Atualmente a regressão por mínimos quadrados parciais tornou-se uma ferramenta padrão para modelagem de relações lineares entre medições multivariadas.O PLS é eficaz para modelar regressões com múltiplas variáveis resposta, não é afetado por multicolinearidade e produz fatores que tenham alto poder de predição (MORELLATO, 2010).
A regressão PLS é chamada de "Mínimos Quadrados Parciais" (PartialLeastSquares) porque os parâmetros são estimados por uma série de regressões de mínimos quadrados, enquanto o termo "parciais" decorre do procedimento de estimação iterativa dos parâmetros em blocos (por variável latente) em detrimento de todo o modelo, simultaneamente (Lee et al., 2011).

Departamento de Estatística e Ciências Atuariais – UFS, email: *jacioliveira416@gmail.com*
Departamento de Química – UFS, email: *antoniovilanova10@gmail.com*
Departamento de Química – UFS, email: *nbcj@ufs.br*
Departamento de Estatística – UFRPE, email: *jfilgueiracosta@gmail.com*
Departamento de Estatística – UFRPE, email: *guirocham@gmail.com*
Departamento de Estatística e Ciências Atuariais – UFS, email: *raphael_crad@yahoo.com*

As duas abordagens foram comparadas através da correlação entre os valores estimados e observados e do Erro Quadrático Médio considerando diferentes critérios para definir as partições que formariam conjunto de treinamento e de teste.

## 2. Metodologia

Foi obtido um banco de dados com 96 variáveis observadas em 602 estruturas. O conjunto de dados foi organizado de forma a verificar a presença de observações faltantes ou possíveis erros de digitação e remover quaisquer variáveis qualitativas para possibilitar a aplicação das técnicas de PLS e PCR. Foram analisadas as correlações das variáveis independentes entre si e com as variáveis dependentes em busca de um direcionamento sobre as possíveis variáveis mais provavelmente relevantes, e também uma inspeção visual através de gráficos de dispersão. Apesar da presença de *outliers*, as técnicas foram utilizadas antes de sua remoção para avaliar como a exclusão de valores extremos pode afetar o desempenho dos métodos.

A seguir estão listados os passos para a construção dos modelos considerados:
1- Divisão do banco de dados em conjuntos de treino e teste;
2- Aplicação da técnica de PLS no conjunto de treino e geração de índices de importância das variáveis;
3- Predição da variável resposta *y* para o conjunto de treino e eliminação das variáveis irrelevantes e ruidosas;
4- Construção de um gráfico para identificação do melhor subconjunto de variáveis e validação das variáveis selecionadas no conjunto de teste;
5- Comparação do desempenho dos diferentes índices frente ao método Stepwise.

O mesmo procedimento foi realizado com PCR para que fosse possível comparar os diferentes modelos através do Erro Quadrático Médio. A interferência na qualidade do ajuste decorrente da modificação dos conjuntos de treinamento e teste também foi avaliada.

Para construir os modelos foi utilizada também análise de agrupamento, técnica que identifica a similaridade entre casos ou variáveis dividindo em grupos e em seguida foi avaliado se o modelo obtido a partir de determinados grupos seria bom para fazer previsão no restante dos dados.

## 3. Resultados e discussões

O banco de dados estava organizado de forma que 93 variáveis eram explicativas, observadas com o objetivo de identificar quais delas ajudam a explicar o comportamento dastrês variáveis restantes: o $IC_{50}$ e suas transformações $\ln(IC_{50})$ e $1/IC_{50}$. A Tabela 1 apresenta média e desvio padrão das variáveis $IC_{50}$, $\ln(IC_{50})$ e $1/IC_{50}$.

Tabela 1: Média e desvio padrão das variáveis resposta.

| **Estatística** | **$IC_{50}$** | **$\ln(IC_{50})$** | **$1/IC_{50}$** |
|---|---|---|---|
| Média | 71,743 | -1,120 | 40,993 |
| D.Padrão | 24,815 | 2,824 | 131,888 |

Fonte: próprio autor

Dividiu-se o banco de dados em duas matrizes, *X* e *Y*, sendo que em *X* estão contidas as 93 variáveis explicativas e em *Y* as 3 variáveis resposta e, em seguida, os dados foram padronizados, já que a

ordem de grandeza era bem diferente entre as variáveis e, se as escalas originais fossem mantidas, algumas delas seriam responsáveis por grande parte da variabilidade simplesmente por assumirem valores num intervalo muito maior.

Ao tentar ajustar um modelo com todas as observações e observar que pouca variabilidade foi explicada, surgiu a hipótese de que o efeito de algumas variáveis explicativas se manifeste de modo diferente dependendo do tipo de estrutura. Portanto o passo seguinte foi particionar o conjunto original em grupos dentro dos quais as estruturas tenham um comportamento semelhante e verificar em cada grupo como as variáveis se relacionam.

Utilizando a técnica de agrupamento K-means, as observações foram separadas em "clusters" (grupos homogêneos de observações dos dados, identificadossegundo alguma distância estatística) que foram usados como base para construir novamente os modelos de regressão múltipla PLS e PCR. A análise de agrupamentos identificou 4 clusters nos dados em estudo, sendo assim foram ajustados modelos PLS e PCR testando cada um dos clusters.
A tabela 2 mostra o resumo dos modelos PCR e PLS levando em consideração 10 componentes principais. O modelo A é o ajuste com todas as observações dos dados, os modelos B, C, D e E foram ajustados com o cluster 1, cluster 2, cluster 3 e cluster 4 respectivamente. Já o modelo F foi ajustado com as observações extremas (maiores e menores valores de $IC_{50}$) dos dados numa tentativa de identificar se o padrão de comportamento das variáveis explicativas muda muito entre as estruturas com melhor e pior desempenho para a variável dependente de interesse e possivelmente permitir mais facilmente a visualização das relações entre elas.

Tabela 2: Comparação entre os modelos PCR e PLS para diferentes conjuntos de treinamento e teste utilizando o percentual da variabilidade explicada para cada variável dependente no conjunto de treinamento, a correlação entre as previsões para a variável melhor descrita no conjunto de teste e o Erro Quadrático Médio.

| **Modelo** | **$IC_{50}$** | **ln($IC_{50}$)** | **1/$IC_{50}$** | **Correlação entre previstos e originais** | **EQM** |
|---|---|---|---|---|---|
| A (PCR) | 14,84% | 43,01% | 7,26% | 0,6558 | 1,0959 |
| A (PLS) | 29,76% | 55,09% | 15,28% | 0,7422 | 1,229 |
| B (PCR) | 36,39% | 48,72% | 14,93% | 0,5973 | 1,2807 |
| B (PLS) | 41,55% | 68,28% | 27,79% | 0,7069 | 1,3839 |
| C (PCR) | 48,32% | 55,11% | 12,05% | 0,0910 | 1,4662 |
| C (PLS) | 81,64% | 59,56% | 14,61% | 0,2211 | 3,2241 |
| D (PCR) | 56,65% | 44,90% | 6,66% | 0,2536 | 1,3397 |
| D (PLS) | 71,37% | 61,32% | 19,06% | 0,2431 | 1,6494 |
| E (PCR) | 47,45% | 51,91% | 50,77% | 0,1635 | 2,0763 |
| E (PLS) | 75,32% | 69,29% | 59,06% | 0,2688 | 1,3894 |
| F (PCR) | 27,91% | 70,97% | 17,22% | 0,6211 | 1,6828 |
| F (PLS) | 54,16% | 84,25% | 26,20% | 0,6759 | 1,7688 |

Fonte: próprio autor

As colunas 2, 3 e 4 mostram a quantidade de variabilidade das variáveis resposta $IC_{50}$, $\ln(IC_{50})$ e $1/IC_{50}$ explicada por cada modelo. A quinta coluna mostra a correlação dos valores previstos para o conjunto teste com os originais da variável que cada modelo mais consegue explicar e a última coluna traz o Erro Quadrático Médio (EQM) de cada modelo.

Pode-se observar na tabela acima que os modelos PLS foram melhores tanto para explicar a variabilidade dos dados quanto para fazer previsão dos mesmos, no entanto os modelos com menor erro quadrático médio foram os modelos utilizando a técnica PCR, com exceção do modelo E. Como o cálculo do EQM é sensível à presença de valores muito discrepantes, é possível que os modelos PCR, afim de melhor descrever essas observações, tenham prejudicado a descrição de pontos mais próximos da média, enquanto a regressão PLS consegue estimativas um pouco melhores para a maioria dos pontos negligenciando observações muito afastadas.O modelo A (tanto PLS como PCR) com todas as observações foi bom para prever os dados, mas não pode ser considerado bom para explicar a variabilidade, porém é preciso destacar que especificamente nesse caso o conjunto de treinamento e teste são iguais.

Os modelosB e F, utilizando a técnica PLS, que foram ajustados a partir do primeiro agrupamento e dos pontos extremos, respectivamente,foram consideradossatisfatórios, pois explicam bem a variabilidade da variável resposta $\ln(IC_{50})$, conseguem prever bem a mesma e tem erro quadrático médio razoável.

As tabelas 3 e 4 apresentam as médias e desvio padrão das variáveis mais importantes para o modelo ajustado PLS, segundo o índice de importância gerado por cada uma das dez componentes principais.

Tabela 3: Estatísticas das variáveis mais importantes para o modelo.

| **Estatística** | **$VCI_1$** | **$VCI_2$** | **$DHN_1$** | **$DHC_2$** | **$DHC_4$** | **$DHN_7$** | **$DHN_8$** | **$NFDN_1$** | **$NFDC_2$** | **$NFDC_4$** |
|---|---|---|---|---|---|---|---|---|---|---|
| Média | 7,983 | 5,932 | 0,093 | 0,034 | 0,024 | 0,064 | 0,044 | 0,056 | 0,146 | 0,172 |
| D.Padrão | 1,342 | 1,007 | 0,067 | 0,022 | 0,016 | 0,047 | 0,033 | 0,029 | 0,066 | 0,083 |

Fonte: próprio autor

Tabela 4: Estatísticas das variáveis mais importantes para o modelo.

| **Estatística** | **$NFD_6$** | **$RFDC_2$** | **$RFDN_7$** | **$NSC_2$** | **$RSC_4$** | **$ESC_2$** | **$ESC_4$** | **$DLC_2$** | **$DLC_6$** | **$DLN_7$** |
|---|---|---|---|---|---|---|---|---|---|---|
| Média | 0,112 | 0,104 | 0,074 | 1,153 | 0,218 | 0,460 | 0,424 | 0,074 | 0,018 | 0,014 |
| D.Padrão | 0,064 | 0,034 | 0,028 | 0,120 | 0,004 | 0,036 | 0,037 | 0,079 | 0,047 | 0,014 |

Fonte: próprio autor

## 4. Conclusões

No geral as técnicas multivariadas PLS e PCR utilizadas tiveram um bom desempenho, pois permitiram identificar um modelo que se ajustasse bem ao conjunto teste e que conseguiu descrever bem os dados, porém a técnica PLS se mostrou mais robusta nesse estudo com relação à capacidade de previsão.

Com base no índice de importância gerado por cada um dos dez componentes principais foi possível identificar as variáveis mais importantes para o modelo de regressão PLS gerado, são elas: $VCI_1$, $VCI_2$, $DHN_1$, $DHC_2$, $DHC_4$, $DHN_7$, $DHN_8$, $NFDN_1$, $NFDC_2$, $NFDC_4$, $NFD_6$, $RFDC_2$, $RFDN_7$, $NSC_2$, $RSC_4$, $ESC_2$, $ESC_4$, $DLC_2$, $DLC_6$ e $DLN_7$.

## ReferênciasBibliográficas

ANDERSON, T.W. **An Introduction to Multivariate StatisticalAnalisys**. Third edition.Stanford,CA:Wiley-Interscience, 1984.

GELADI,P; Kowalski, B.R.,**Partial Last Squares: A Tutorial**, Elsevier Science Publishers B.V.,Amsterdã,1986.

GOMES, A.A.**Algoritmos das Projeções Sucessivas Aplicado a Seleção de Variáveis em Regressão PLS**.2012.121f. Dissertação de Mestrado-Universidade Federal da Paraíba, Pernambuco,2102.

MANLY, B.J.F.**Métodos Estatísticos Multivariados: Uma Introdução**.3 ed. Porto Alegre:Bookman,2008.

MORELLATO, S.A.**Modelos de Regressão PLS com Erros Heterocedásticos.** 2010.60f.Dissertação de Mestrado-Universidade Federal de São Carlos, São Paulo,2010

MORGANO, M. A. et al. **Determinação de Proteína em Produção de café Cru por Espectropia NIR e PLS**, Ciência e Tecnologia de Alimento, Campinas, dezembro 2004. Disponível em: http://www.redalyc.org/comocitar.oa?id=395940073005; acesso em:19-01-2018

SAAD, D. S. **Aplicação de Técnicas Estatísticas Multivariadas em Dados de Cerâmica Vermelha Produzida no Rio Grande do Sul**. 2009.166f. Dissertação de Mestrado-Universidade Federal de Santa Maria, Rio Grande do Sul, 2009.

WOLD, S.;SJÖSTRÖME, M.; ERIKSSON, L. **PLS-Regression: A Basic Tool of Chemometrics**, Elsevier, Chemometrics and Intelligent Laboratory Systems, v.58, p.109-130, 2001.