ISSN: 2317-0840

Estudo de técnicas multivariadas para seleção de variáveis em grandes bancos de dados: uma aplicação envolvendo dados de inibição (IC_{50})

Jaciele J. Oliveira[1†], Antônio Luiz S. V. Costa[1], João B. F. Costa[2], Guilherme R. Moreira[3], Nivan B. Costa[1], Carlos R. A. Daniel[1]

[1]*Departamento de Estatística e Ciências Atuariais, Universidade Federal de Sergipe.*
[2]*Mestrando em Biometria e Estatística Aplicada, Universidade Federal Rural de Pernambuco.*
E-mail: jfilgueiracosta@gmail.com.
[3]*Prof. Biometria e Estatística Aplicada, Universidade Federal de Pernambuco.*
E-mail: guirocham@gmail.com.

Resumo: *A análise multivariada é um meio eficiente na análise de grandes bancos de dados contendo inúmeras variáveis, pois tais técnicas podem ser utilizadas para obter um número reduzido de variáveis sem perda de informação útil. Este trabalho tem por objetivo estudar as técnicas de regressão múltipla, Regressão Por Componentes Principais e Mínimos Quadrados Parciais em problemas de seleção de variáveis e avaliar o desempenho destas estratégias em um banco de dados real. O banco de dados utilizado apresenta 602 estruturas e 93 variáveis buscando descrever o comportamento das variáveis resposta Índice de Concentração (IC_{50}) e suas transformações ln(IC_{50}) e 1/ IC_{50}. O IC_{50} é uma medida da potência de uma substância no processo de inibição de uma função química ou biológica, indicando quanto da substância é necessária para inibir um dado processo pela metade, portanto, quanto menor o IC_{50}, mais ativo é o composto. Foram ajustados modelos particionando os dados em conjuntos de treinamento e teste. Os dados também foram submetidos a uma análise de agrupamento numa tentativa de separar grupos de compostos semelhantes entre si. A presença de outliers e sua influência nos ajustes foram avaliadas. No geral as técnicas utilizadas tiveram um desempenho satisfatório comparando valores de erro quadrático médio, permitindo identificar um modelo que se ajustou bem ao conjunto teste e conseguiu descrever bem os dados. Na maioria dos casos, a técnica Mínimos Quadrados Parciais, apresentou melhores resultados nesse estudo. Por fim, foi possível destacar as 20 variáveis mais relevantes para o modelo.*

Palavras-chave: Regressão Por Componentes Principais; Mínimos Quadrados Parciais; Índice de Concentração.

Abstract: *Multivariate analysis is an efficient way to analyze large databases containing numerous variables, since such techniques can be used to obtain a small number of variables without loss of useful information. This work aims to study the multiple regression techniques Principal Components Regression and Partial Least Squares in variables selection problems and to evaluate the performance of these strategies in a real database. The database used presents 602 structures and 93 variables seeking to describe the behavior of the response variables Concentration Index (IC_{50}) and its ln(IC_{50}) and 1 / IC_{50} transformations. The IC_{50} is a measure of the potency of a substance in the process of inhibiting a chemical or biological function, indicating how much of the*

[†]Autora correspondente: jacioliveira416@gmail.com.

substance is needed to inhibit a given process in half, so the lower the IC_{50}, the more active the compound is. Models were adjusted by partitioning the data into training and test sets. The data were also subjected to a cluster analysis in an attempt to separate groups of similar compounds from each other. The presence of outliers and their influence on adjustments were assessed. In general, the techniques used had a satisfactory performance comparing mean square error values, allowing to identify a model that fit well with the test set and was able to describe the data well. In most cases, the Partial Least Squares technique showed better results in this study. Finally, it was possible to highlight the 20 most relevant variables for the model.

Keywords: Partial Least Squares; Principal Components Regression; Concentration Index.

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 de complexidade 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 três métodos de análise multivariada de dados: Regressão por Componentes Principais (PCR – *Principal Components Regression*), Mínimos Quadrados Parciais (PLS – *Partial Least Squares*) 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 quanto maior 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; SAPORTA, 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" (*Partial Least Squares*) 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).

Metodologia

Foi obtido um banco de dados com 96 variáveis observadas em 602 observações. 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* identificada nos gráficos de dispersão, as técnicas foram utilizadas antes de sua remoção para avaliar como a exclusão de valores extremos pode afetar o desempenho dos métodos.

O banco de dados em estudo foi avaliado usando PLS e PCR, considerando modelos para descrever as variáveis resposta O percentual da variabilidade explicada à medida que novas componentes são incorporadas ao ajuste foi registrado até a inclusão de 10 componentes para cada caso.

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.

A seguir estão listados os passos para a construção dos modelos considerados:

1. Divisão do banco de dados em conjuntos de treino e teste. Essa divisão foi feita considerando as 602 observações, para isso foi realizada a análise de agrupamentos K-means que resultou em 4 clusters e outro grupo foi formado pelas respectivos valores extremos de IC_{50} (considerando como regra os pontos menores que 0,02 e maiores que 10,0), cada um desses grupos é um conjunto treino;
2. Aplicação da técnica de modelagem no conjunto de treino e geração de índices de importância das variáveis, obtendo assim o modelo ajustado;
3. Utilizando o modelo ajustado na predição da variável resposta *y* para o conjunto de treino e predição também no conjunto teste.
4. Construção de gráficos de dispersão entre valores ajustados e valores observados para identificação do melhor subconjunto de observações e validação das variáveis selecionadas no conjunto de teste;
5. Escolha do melhor modelo e geração de índices de importância através dos componentes principais para identificar as variáveis mais relevantes neste modelo.

O procedimento foi realizado tanto com PCR como PLS para que fosse possível comparar os diferentes modelos 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. 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. A técnica de agrupamento utilizada foi o agrupamento *K-means*, o qual utiliza a distância euclidiana para medir a distância entre as observações. Para identificar quantos grupos seriam necessários para separar os dados foi feito o gráfico dendrograma, o qual permite agrupa estruturas

semelhantes a partir das distâncias entre elas, assim percebeu-se que quatro grupos seriam bons para agrupar os dados.

Em primeiro plano foi ajustado um modelo PLS com todas as variáveis e observações dos dados, o qual explicou muito pouco as variáveis respostas, em seguida foram ajustados modelos com cada um dos clusters formado na análise de agrupamentos e por último ajustou um modelo considerando somente as observações referentes a valores extremos da variável resposta.

As respectivas técnicas de análise multivariada escolhida neste estudo foram realizadas no software estatístico R (R CORE TEAM, 2020), utilizando pacotes específicos para a análise como o `pls`.

As principais funções para ajuste de modelos são `pcr` e `plsr`. Na sua forma mais simples, a chamada de função para ajustar modelos é `plsr(formula, ncomp, dados)`, onde o `plsr` pode ser substituído por `pcr` ou `mvr`. O argumento `formula` é alguma relação do tipo $Y \sim X$, em que Y é um conjunto de variáveis dependentes e X é um conjunto de variáveis explicativas, `ncomp` é o número de componentes que se deseja incluir, e `dados` é o data frame que contém as variáveis a serem usadas no modelo. A função retorna um modelo que pode ser usado para prever novas observações.

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 das três 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 $1/IC_{50.}$ O IC50 é uma medida da potência de uma substância no processo de inibição de uma função química ou biológica, indicando quanto da substância é necessária para inibir um dado processo pela metade, portanto, quanto menor o IC50, mais ativo é um composto.

Tabela: Média e desvio padrão das variáveis respostas

Estatística	IC50tg	lnIC50	1/IC50
Média	71.743	-1.120	40.9929
D.Padrão	24.8149	2.8244	131.8875

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 molecular dos fármacos. 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 dependentes e independentes se relacionam.

Utilizando a técnica de agrupamento *K-means* (MACQUEEN, 1967), as observações (linhas da matriz de dados) foram agrupadas em "clusters" (grupos homogêneos de observações dos dados, identificados segundo alguma distância estatística, neste caso a distância euclidiana) 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 utilizando cada um dos clusters (por exemplo um modelo PCR e um modelo PCR ajustado com os componentes do cluster 1).

A Tabela 2 mostra o resumo dos modelos PCR e PLS levando em consideração apenas os 10 primeiros componentes principais, pois eles explicam mais de 98% da variabilidade dos dados, ou seja, os outros componentes explicam minimamente a variabilidade. 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 apenas com as observações extremas de IC_{50} (considerando como regra os pontos menores que 0,02 e maiores que 10,0) 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

As colunas 2, 3 e 4 da Tabela 2 apresentam a quantidade de variabilidade das variáveis resposta IC_{50}, $\ln(IC_{50})$ e $1/IC_{50}$ explicada por cada modelo. A quinta coluna apresenta a correlação dos valores previstos para a variável resposta no conjunto teste com os valores originais da variável resposta que cada modelo mais consegue explicar (seja o IC_{50} ou suas transformações $\ln(IC_{50})$ e $1/IC_{50}$), ou seja, a maior correlação observada entre valores previstos e ajustados para cada variável resposta no modelo especificado na primeira coluna 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, pois observamos nas colunas 2, 3, 4 e 5 que o modelos PLS apresentam maiores valores quando comparados as modelos PCR, 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, a fim 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 dado que a maior porcentagem de variabilidade explicada apresentada é de 55,09%(um valor baixo), porém é preciso destacar que especificamente nesse caso o conjunto de treinamento e teste são iguais.

Ao comparar as técnicas PCR e PLS usando dados de misturas químicas complexas simulados, Wentzell e Montoto encontraram resultados diferentes dos apresentados aqui, não obtiveram diferenças significativas nos erros de predições de ambos os modelos, apesar de na maioria das vezes PLS exigir menos variáveis latentes do que PCR, mas isso não influenciou a capacidade preditiva. Yaroshchyk *et al*. Também encontraram resultados semelhantes aos de Wentzell e Montoto utilizando as mesmas técnicas em dados para a análise quantitativa do teor de Fe no minério de ferro medido usando Espectroscopia de decomposição induzida por laser.

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

- DHN-A desidrina é uma família múltipla de proteínas presentes nas plantas, que é produzida em resposta ao frio e à seca.
- NFDN- nifedipina é um fármaco bloqueador dos canais de cálcio utilizado como hipotensor, vasodilatador e tocolítico
- DLC- carbono amorfo hidrogenado- variedade alotrópica do carbono que se apresenta na forma hidrogenada.

Tabela 3: Estatísticas das variáveis mais importantes para o modelo B (PLS)

Estatística	VCI_1	VCI_2	DHN_1	DHC_2	DHC_4	DHN_7	DHN_8	$NFDN_1$	$NFDC_2$	$NFDC_4$
Média	7.9833	5.9324	0.0931	0.0341	0.0236	0.0641	0.0436	0.0564	0.1457	0.1724
D.Padrão	1.342163	1.0067	0.0669	0.0222	0.0159	0.0467	0.0326	0.0289	0.0658	0.0825

Tabela 4: Estatísticas das variáveis mais importantes para o modelo F (PLS)

Estatística	NFD_6	$RFDC_2$	$RFDN_7$	NSC_2	RSC_4	ESC_2	ESC_4	DLC_2	DLC_6	DLN_7
Média	0.1115	0.1043	0.0741	1.1528	0.2177	0.4601	0.4244	0.0741	0.0183	0.0142
D.Padrão	0.0635	0.0342	0.0278	0.1201	0.0044	0.0361	0.0374	0.0786	0.0472	0.0138

As tabelas 3 e 4 apresentam as médias e desvios padrão das variáveis mais importantes para o modelo ajustado PLS, modelo B e modelo F respectivamente, segundo o índice de importância gerado por cada um dos dez componentes principais. Essas variáveis representam propriedades dos potenciais fármacos, como a composição química, o número de átomos, distâncias e ângulos entre eles na estrutura molecular. Por exemplo: o DHN (desidrina) é uma família múltipla de proteínas presentes nas plantas, que é produzida em resposta ao frio e à seca, O NFDC (nifedipina) é um fármaco bloqueador dos canais de cálcio utilizado como hipotensor, vasodilatador e tocolítico e o DLC (carbono amorfo hidrogenado) é uma variedade alotrópica do carbono que se apresenta na forma hidrogenada.

Conclusões

Os modelos ajustados com todas as observações não explicaram bem a variabilidade da variável resposta IC_{50} e suas transformações. Ao ajustar os modelos com os 4 clusters identificados na análise de agrupamentos, observamos que apenas o ajuste com o cluster 1 apresentou resultados de explicação da variável resposta e predição razoáveis. O modelo ajustado com as observações referentes aos valores extremos do IC_{50} explicou bem o $\ln(IC_{50})$ e apresentou boa predição no conjunto teste.

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.

Agradecimentos

O presente trabalho teve o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – BRASIL.

Referências Bibliográficas

BOROUCHE, J. M., SAPORTA. G. *Análise de dados*. Zahar Editores. Rio de Janeiro, 1982.

MACQUEEN, J. B. *Some Methods for Classification and Analysis of Multivariate Observations*. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California, 1967.

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

R CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 2000. Disponível em: https://www.R-project.org/.

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.

YAROSHCHYK, P; DEATH, D.L. e SPENCER S.J. *Comparison of principal components regression, partial least squares regression, multi-block partial least squares regression, and serial partial least squares regression algorithms for the analysis of Fe in iron ore using LIBS.* Journal Analytical Atomic Spectrometry, 1 ed, 2012.

WENTZELL, P.D. e MONTOTO L.V. *Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures*. Chemometrics and Intelligent Laboratory Systems, V. 65, 2003, Pages 257-279.

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