Análise multivariada de dados genômicos simulados a partir do projeto TCGA

Multivariate analysis of simulated genomic data from the TCGA project

Heitor Baldo*
Análise Multivariada (MAE5776)

1º Semestre, 2020

Resumo

In this work, we analyse gene and protein expression databases simulated using the OmicsSIMLA tool from the TCGA project. Various multivariate statistical techniques were performed on the protein expression database, such as principal component analysis for dimensionality reduction, linear discriminant analysis, and cluster analysis for group classification. At the end of the study, an integration of both databases was carried out using the partial least squares regression model and canonical correspondence analysis, which showed that the two databases are poorly correlated.

Conteúdo

*NUSP: 11571801 / hbaldo@usp.br

1 Introdução

A análise multivariada lida com dados contendo observações em duas ou mais variáveis, cada uma medida em um conjunto de indivíduos, e com métodos estatísticos que analisam simultaneamente estas variáveis e indivíduos. Convenciona-se arranjar dados multivariados em uma "matriz de dados", $Y_{n\times p}$, sendo n o número de indivíduos e p o número de variáveis [7].

Dentre as diversas técnicas estatísticas multivariadas existentes na literatura, podemos destacar a Análise de Componentes Principais (do inglês *Principal Components Analysis* (PCA)), Escalonamento Multidimencional (ou Análise de Coordenadas Principais (PCoA)), Análise Fatorial (FA), Análise de Correnpondência, Análise Discriminante Linear de Fisher (LDA), Análise de Correspondência Canônica (CCA), Mínimos Quadrados Parciais (PLS) e Análise de Agrupamentos (*Clustering Analysis* (CA)). Elas podem ser utilizadas para redução de dimensionalidade, integração de bancos de dados, etc. O leitor interessado em aprofundar seus estudos em cada uma das técnicas precedentes, pode consultar os livros-texto [1], [4], [6] e [7] para abordagens mais teóricas e [3] e [5] para abordagens mais práticas (especialmente práticas na linguagem de programação R).

O foco principal deste texto é empregrar diversas dessas técnicas multivariadas em dados genômicos simulados a partir do projeto *The Cancer Genome Atlas* (TCGA) utilizando a ferramenta de simulação OmicsSIMLA[1] [2].

2 Materiais e Metodologia

2.1 Dados de expressões gênicas e proteicas

A partir do projeto TCGA, utilizou-se a ferramenta OmicsSIMLA para simular dados de câncer de ovário que foram postos em cinco bancos de dados, cada um contendo informações genômicas específicas: CNV (*Copy Number Variation*), Metilação (Methy), Expressão Gênica (Exp), Expressão Gência Normalizada (NorExp)) e Expressão Proteica (Protein).

As unidades amostrais correspondem a pacientes com tempo de sobrevivência inferior a 3 anos (casos – outcome = 1) e superior a 3 anos (controle – outcome = 0). Os dados foram gerados supondo a presença de um loco cromossômico (eQTM) com nível de metilação diferente para casos e controles, o qual influencia os genes LRIG1, TCEAL8 e MARCH9. A simulação dos casos e controles foi feita condicionalmente à expressão gênica destes três genes e do LRRN4. Foram simulados dados para 1000 pacientes, sendo 500 casos e 500 controles.

Para esses cincos bancos, foram produzidas 100 réplicas. Este texto embasa-se na análise da réplica número *quatro* dos bancos de dados de Expressão Gênica Normalizada (NorExp), que são dados da intensidade de expressão gênica normalizados, em que foi eliminado ruídos aleatórios, e no banco de Expressão Proteica (Protein), que são dados da expressão proteica normalizada para os mesmos genes do banco de expressão gênica.

A partir desses dois bancos de dados, foi realizada uma análise de significância das variáveis através de ANOVAs e na subsequente construção de um *Volcano Plot*, a partir do qual foi selecionado as 99 variáveis mais expressivas de NorExp e as 89 mais expressivas de Protein.

Tendo realizado esta seleção, obtemos duas matrizes de dados, $Y_{1000\times 99}$ para NorExp e $Y_{1000\times 89}$ para Protein. Na matriz $Y_{1000\times 89}$ realizamos as seguintes análises multivariadas: Análise de Componentes Principais, Análise Discriminante Linear de Fisher e Análise de Agrupamentos. Na sequência, realizamos uma integração dos bancos de dados $Y_{1000\times 89}$ e $Y_{1000\times 99}$ primeiramente

[1] https://omicssimla.sourceforge.io

através da técnica de Minímos Quadrados parciais, seguida de uma integração por Análise de Correspondência Canônica.

Na seção abaixo, apresentamos um brevíssimo resumo teórico sobre cada uma das análises aplicadas.

2.2 Análise de Componentes Principais (PCA)

A questão fundamental que a *Análise de Componentes Principais* procura responder é: dada uma matriz de dados $Y_{n\times p}$, como podemos combinar as p variáveis para obtermos $p = 2$ ou $p = 1$ variáveis? Ou seja, é um problema de *redução de dimensionalidade*.

A decomposição espectral da matriz de covariância, $\Sigma_{p\times p} = P_{p\times p}\Lambda_{p\times p}P'_{p\times p}$, permite uma representação dos dados em eixos ortogonais e nas direções de máxima variação (total) dos dados. Tais eixos são chamados *componentes principais* (CP). Pela aproximação $\Sigma = \sum_{j=1}^{p}\lambda_j P_j P_j \approx \sum_{j=1}^{m}\lambda_j P_j P_j$, os CP, Z_i, podem ser expressos por

$$Z_i = (Z_{i1}, ..., Z_{im})' = P'_{m\times p}Y_{i_{p\times 1}} \in \mathbb{R}^m,$$

$$Y_i \in \mathbb{R}^p \to Z_i = P'_{m\times p}Y_{i_{p\times 1}} \in \mathbb{R}^m; \quad Z_{ki} = P'_k Y_{p\times 1}; \quad Var(Z_{ki}) = \lambda_k,$$

$\lambda_1 \geq ... \geq \lambda_p$ autovalores. Em geral, retemos um número de CP na análise que preserve "grande" parte da variância total dos dados, ou seja, queremos garantir que "grande" parte da variabilidade de cada variável original seja explicada pelos CP.

2.3 Análise Discriminante Linear de Fisher (LDA)

A análise discriminante é uma técnica da estatística multivariada utilizada para discriminar e classificar objetos. A classificação ou alocação pode ser definida como um conjunto de regras que serão usadas para alocar novos objetos. Em outras palavras, considerando G grupos, $\pi_1, ..., \pi_G$, o objetivo da análise discriminante é, essencialmente, alocar os indivíduos em cada um desses grupos, baseado em algum critério de classificação.

Na *Análise Discriminante Linear de Fisher*, queremos encontrar a direção discriminante ótima que maximiza B (eixo de variação entre grupos) relativamente a W (eixo de variação dentro de grupos)[2]. A partir da decomposição espectral da matriz $W^{-1}B$, queremos encontrar um vetor a que resolva o problema de otimização

$$max_a \frac{a'Ba}{a'Wa}.$$

Dada uma observação x e o vetor a que maximiza a razão anterior, a função linear $a'x$ é denominada de *função discriminante linear de Fisher*. O vetor a na função discriminante linear de Fisher é o autovetor de $W^{-1}B$ correspondente ao maior autovalor. Uma vez que a função discriminante linear foi calculada, uma observação x pode ser alocada em um dos G grupos com base no seu escore discriminante $a'x$.

[2]Em que $W = S_W$ é a matriz de soma de quadrados e produtos cruzados dentro de grupo e $B = S_B$ é a matriz de soma de quadrados e produtos cruzados entre de grupos.

2.4 Análise de Agrupamentos (CA)

A *Análise de Agrupamentos* objetiva a formação de grupos de unidades amostrais (agrupamento de observações) em grupos homegêneos internamente e heterogêneos externamente, podendo também identificar similaridades entre variáveis (agrupamento de variáveis). Em outras palavras, é a classificação de objetos em diferentes grupos, cada um dos quais deve conter os objetos semelhantes segundo alguma função de distância estatística.

A análise de agrupamentos engloba diversos métodos, que podem ser divididos em *métodos de agrupamentos hierárquicos* e *métodos de agrupamentos não-hierárquicos* (ou *métodos de partição*):

Métodos Hierárquicos Aglomerativos: os agrupamentos hierárquicos partem dos objetos individuais (n) para a formação de um único grupo. Dentre os métodos desse tipo, podemos citar:

- Método do Vizinho mais Próximo/Perto (Ligação Simples);
- Método do Vizinho mais Distante/Longe (Ligação Completa);
- Método das Médias das Distâncias (Ligação Média);
- Método da Centróide;
- Método de Ward.

Métodos de Partição: os agrupamentos não-hierárquicos buscam a partição de n objetos em K grupos. Dentre os métodos desse tipo, podemos citar:

- Algoritmo das K-Médias.

2.5 Mínimos Quadrados Parciais (PLS)

Dados bancos de dados $X_{n\times q}$ e $Y_{n\times p}$, sendo X uma matriz de variáveis preditoras de Y, o modelo de regressão dos *Mínimos Quadrados Parciais* preocupa-se em obter direções (variáveis Z (os PC de X)) em X que melhor explicam (predizem) Y. Mais ainda, o PLS integra as duas matrizes de dados maximizando a covariância entre elas: encontra vetores reducionistas a de X e b de Y tais que

$$\max_{a\in\mathbb{R}^q; b\in\mathbb{R}^p} \frac{[Cov(a'X; b'Y)]^2}{(a'a)(b'b)}.$$

2.6 Análise de Correspondência Canônica (CCA)

Dados bancos de dados $X_{n\times q}$ e $Y_{n\times p}$, a *Análise de Correspondência Canônica* objetiva encontrar combinações lineares $a'X$ e $b'Y$ tais que elas tenham a maior correlação possível, ou seja, que $Corr(a'X, b'Y)$ seja máxima (a e b são os eixos canônicos que maximizam essa correlação). Tais combinações lineares podem dar *insights* do relacionamento entre os dois conjuntos de variáveis.

A CCA integra duas matrizes de dados maximizando a correlação entre as combinações lineares das matrizes. No entanto, como temos a equivalência

$$[\mathrm{Cov}(a'X; b'Y)]^2 = \mathrm{Var}(a'X)[\mathrm{Corr}(a'X; b'Y)]^2\mathrm{Var}(b'Y).$$

Disso podemos dizer que o PLS é a CCA com regularizações definidas pelos CP de X e de Y. Outra diferença entre a CCA e a PLS é que a CCA é uma análise não-direcionada (simétrica).

3 Resultados e discussão

Nesta seção, iremos interpretar e discutir os resultados obtidos nas análises realizadas. Estaremos interessados em responder questões como: "quais foram as variáveis que mais contribuíram na análise?", "os grupos de pacientes *caso* e *controle* foram bem discriminados na análise?"

3.1 Resultados para o banco de dados Protein

Primeiramente, realizamos uma análise de componentes principais no banco "Protein"para a visualização dos dados no plano (redução de dimensionalidade $p = 89 \rightarrow m = 2$). Sumarizamos os principais resultados nos dois gráficos abaixo.

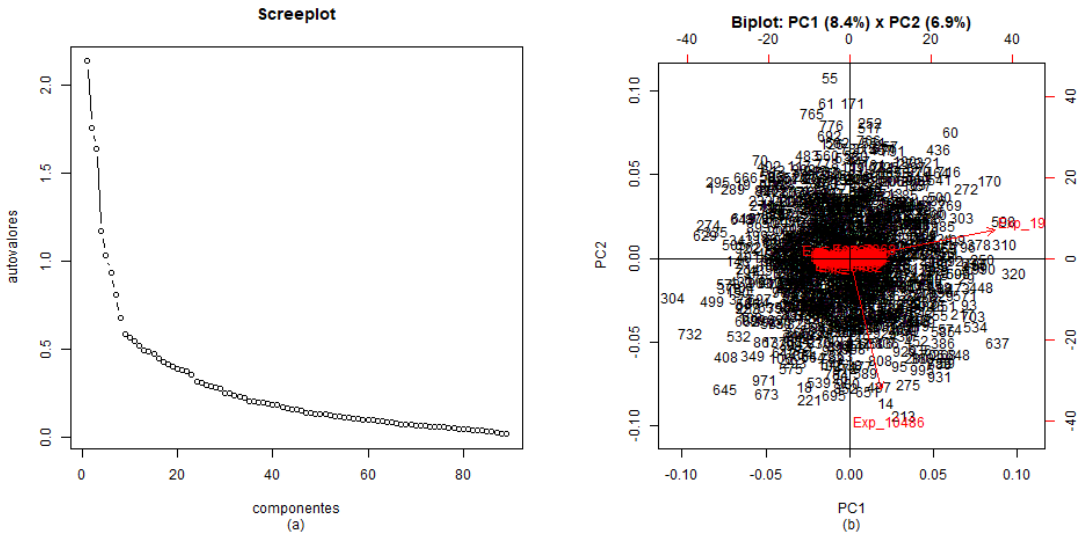

Figura 1: (a) Screeplot: comparação dos autovalores para cada um dos 89 CP. (b) Biplot: representação simultânea dos escores dos CP e dos pesos das variáveis (representação gráfica simultânea das 1000 observações e das 89 variáveis).

O gráfico biplot acima nos informa que a proporção da variância total dos dados explicada pelo CP1 é de aproximadamente 8.4%. Já a proporção da variância total explicada pelo CP2 é de aproxiamdamente 6.9%. Ou seja, juntos, CP1 e CP2, explicam aproximadamente 15.3% da variância total dos dados. Ainda, observamos que uma explicação de 15% da variância dos dados pelos dois primeiros componentes principais não foi suficiente para discriminar os grupos em "caso"e "controle"[3].

Ainda pelo biplot, vemos que a variável Exp_10486 recebe o maior peso (carga) dentre as variáveis, seguida da variável Exp_1922, dominando, assim, a análise.

Em geral, o critério de corte para o screeplot é desconsiderar os autovalores a partir dos quais a variação entre os eles passa a ser pequena (cotovelo do gráfico). Neste caso, escolhemos os dois autovalores localizados na parte mais superior do gráfico, $\lambda_1 = 2.13$ e $\lambda_2 = 1.76$, correpondendo ao CP1 e ao CP2, respectivamente.

[3]Apenas uma observação, pois a finalidade primordial da PCA não é a discriminação de grupos, mas sim a redução de dimensionalidade

Continuando com o banco de dados "Protein", realizamos uma análise discriminante linear de Fisher. As suposições da LDA são: independência das observações, homocedasticidade e prioris iguais. Realizamos o teste M de Box e obtemos p-value $= 0.3353 > 0.05$, portanto não rejeitamos a hipótese de homocedasticidade. Consideramos ainda os prioris estimadas dos dados.

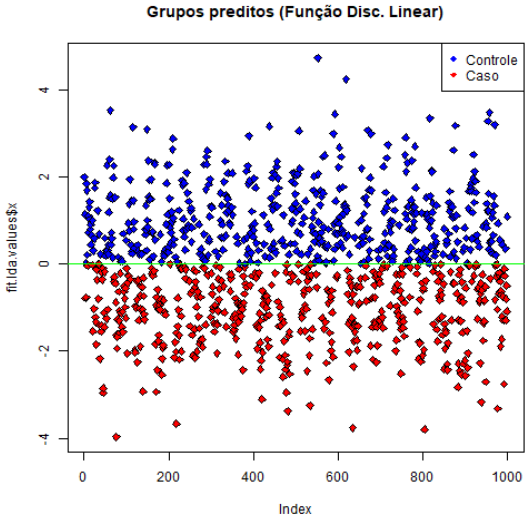

Figura 2: Gráfico de dispersão dos grupos preditos (o eixo de máxima discriminação está representado pela *linha verde*).

	Controle (ME)	Caso (ME)	Controle (CV)	Caso (CV)
Controle	406	92	376	124
Caso	94	408	118	382

Tabela 1: Tabelas de classificação obtidas pelo método empírico (EM) e por validação cruzada (CV).

A partir dos destes dados, observamos que, pelo método empírico (EM), a função discriminante conseguiu classificar corretamente 81.53% das observações no grupo "controle" e 81.27% no grupo "caso" e obtemos uma acurácia de 81.4% na classificação.

Já pelo método de validação cruzada, a função discriminante conseguiu classificar corretamente 75.2% das observações no grupo "controle" e 76.4% no grupo "caso" e obtemos uma acurácia de 75.8% na classificação.

Ademais, obtemos a redução de dimensionalidade: $min(n, p, G-1) = min(1000, 89, 1) = m = 1$ (LD1 é o eixo discriminante).

Finalmente, ainda para o banco "Protein", realizamos uma análise de agrupamento através dos seguintes métodos hierárquico e não-hierárquico: *método do vizinho mais distante (Ligação Completa)* e *K-Médias (algoritmo de Hartigan-Wong)*, respectivamente.

	Contole (LC)	Caso (LC)	Controle (HW)	Caso (HW)
Controle	354	146	304	196
Caso	376	124	343	157

Tabela 2: Tabelas de classificação obtidas pelo método de Ligação Completa (LC) e pelo K-Médias usando o algoritmo de Hartigan-Wong (HW).

A partir destes dados, observamos que o método do vizinho mais distante conseguiu classificar corretamente 47.84% das observações em ambos os grupos, "caso" e "controle", ou seja, obteve uma acurácia de 47.84% na classificação.

Já o algoritmo de Hartigan-Wong conseguiu classificar corretamente 45.94% das observações em ambos os grupos, "caso" e "controle", ou seja, obteve uma acurácia de 45.94% na classificação.

Heatmap

Utilizando o método do vizinho mais distante, construímos um *heatmap* (Figura 4) dos dados. A partir deste heatmap, podemos observar que os genes na parte central têm intensidade de expressão proteica mediana para todos os pacientes (como mostrado no histograma da Figura 4), ou seja, são homogêneos; os três primeiros genes à esquerda têm baixa intesidade de expressão proteica os pacientes da parte inferior, na parte mais central e próximo ao topo do heatmap (vermelho). Constatamos ainda alta intensidade (azul) de expressão proteica do último (à direita) gene (Exp_1922) para os pacientes mais ao topo do heatmap e baixa intensidade para os pacientes na parte inferior.

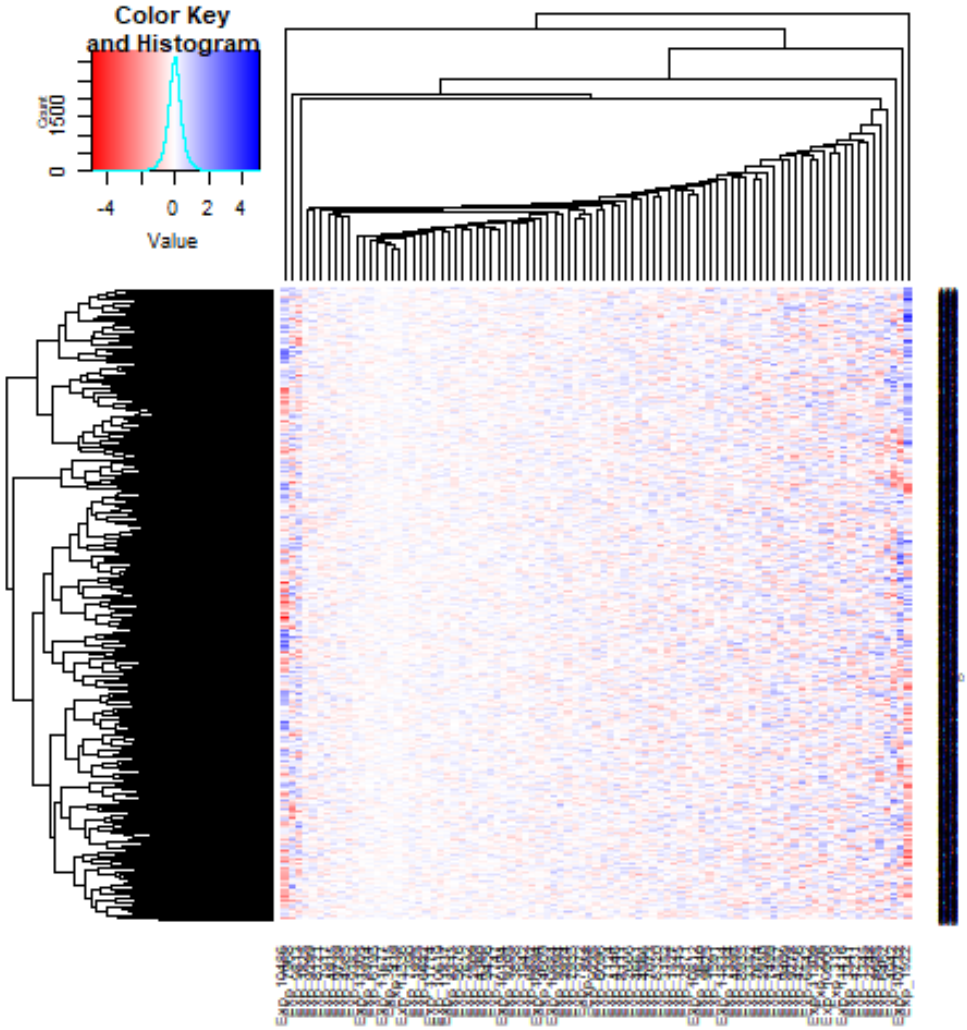

Figura 3: Heatmap construído através do agrupamento hierárquico Ligação Completa (método do vizinho mais distante) com distância euclidiana.

3.2 Resultados da integração dos bancos de dados Protein e NorExp

Vamos analisar agora os dados obtidos da integração dos bancos de dados "Protein" e "NorExp" realizada através do modelo de regressão dos mínimos quadrados parciais e da análise de correspondência canônica.

Para a realização da PLS, consideramos Y = NorExp como matriz de respostas e X = Protein como matriz preditora.

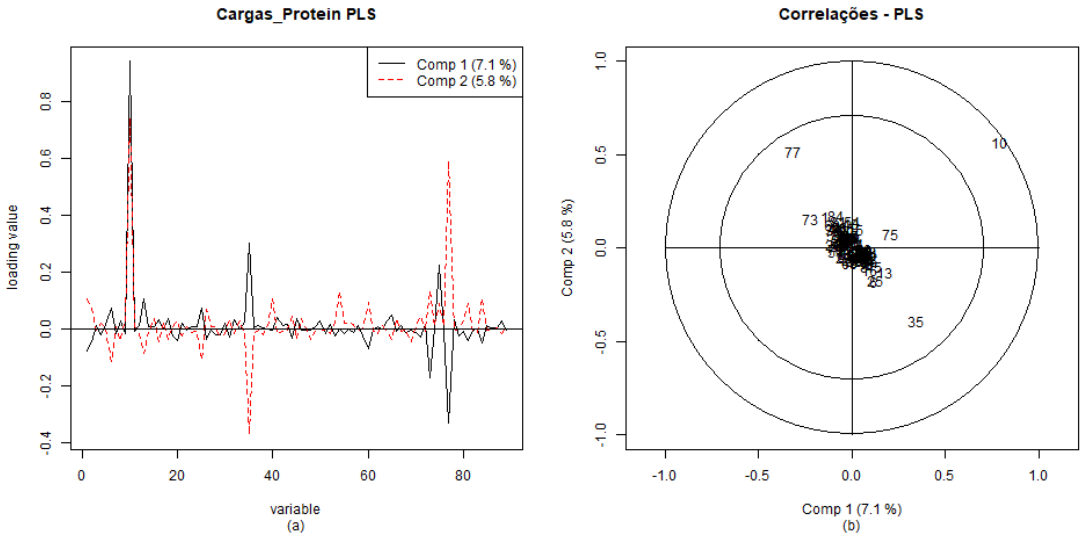

Figura 4: (a) Cargas (coeficienetes) dos dois primeniros componentes de PLS para X. (b) Correlações de cada variável (original) em X com Componentes PLS de X.

O gráfico das cargas (Figura 5 (a)) dos dois primeniros componentes de PLS para X nos informa que o primeiro (Comp1) e o segundo (Comp2) componentes explicam, aproximadamente, respectivamente, 7.1% e 5.8% da variância, explicando em conjunto, aproximadamente, 13% da variância total dos dados.

Pelo gráfico das correlações (Figura 5 (a)), podemos constatar que as variáveis 10 (Exp_1922), 35 (Exp_3716) e 77 (Exp_10732) têm as maiores correlações. No entanto, a maioria das variáveis se concentram no centro do gráfico, indicando que estão pobremente correlacioandas e também que não estão bem representadas por esses componentes PLS.

A Figura 6 (a) apresenta o biplot dos escores e das cargas dos 89 componentes PLS obtidos para X. A partir dele, podemos observar que as variáveis Exp_19 e Exp_10732 recebem os maiores pesos (cargas) dentre as variáveis, dominando, assim, a análise. Notamos ainda que a observação 635 pode ser considerada atípica.

A Figura 6 (b) apresenta o biplot dos escores e das cargas dos 89 componentes PLS obtidos para Y. A partir dele, observamos que as variáveis LRIG e TCEAL8 recebem os maiores pesos (cargas) dentre as variáveis, dominando, assim, a análise.

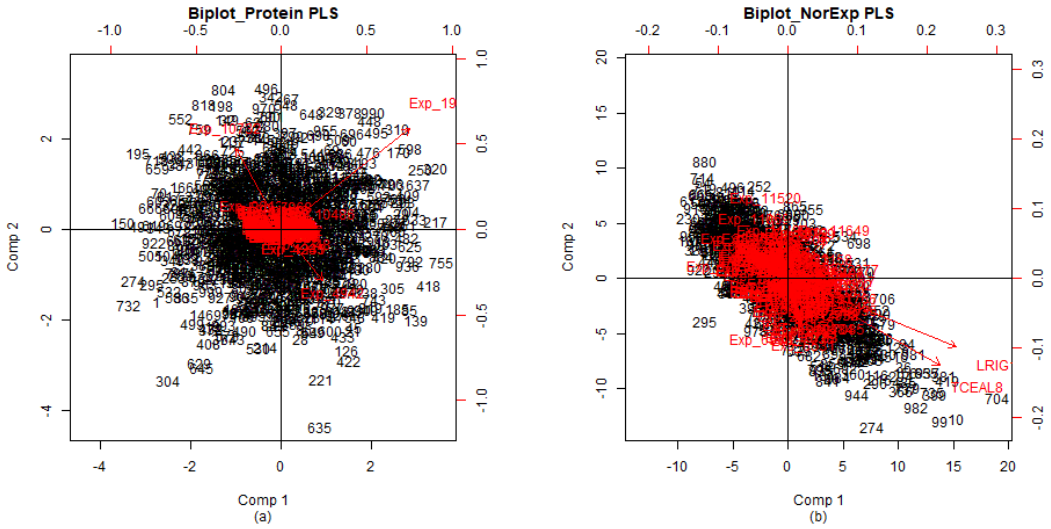

Figura 5: (a) Biplot: representação simultânea dos escores e das cargas dos 89 componentes PLS obtidos para X. (b) Biplot: representação simultânea dos escores e das cargas dos 89 componentes PLS obtidos para Y.

Vamos analisar agora os resultados obtidos da CCA; continuaremos considerando Y = NorExp e X = Protein.

A Figura 7 apresenta os coeficientes (cargas) das variáveis canônicas (U de X e V de Y) obtidos. O cálculo foi realizado utilizando o coeficiente de correlação de Pearson e o maior coeficiente obtido foi igual a 0.5956.

A Figura 8 (a) apresenta o gráfico dos coeficientes das cargas para X e Y. A partir dele, observamos que a maioria das variáveis estão próximas ao centro do gráfico, indicando baixa correlação entre todas elas.

A Figura 8 (b) apresenta o gráfico de dispersão das observações segundo a integração dos bancos de dados. Em relação à correlação, podemos identificar neste gráfico o quão "diferentes" as observações são entre si, por exemplo, a partir dos escores, vemos que as observações 992 e 443 são muito diferentes entre si, por exemplo.

A Figura 9 (a) apresenta um biplot dos escores de X e Y vs cargas de Y. A partir dele, observamos que as variáveis Exp_3615 e Exp_439 recebem os maiores pesos (cargas), porém não podemos decidir sobre sua dominância na análise pois as cargas estão bem distribuídas para as outras variáveis que também recebem cargas semelhantes.

A Figura 9 (b) apresenta um biplot dos escores de X e Y vs cargas de X. A partir dele, vemos que os pesos (cargas) estão bem distribuídos para todas as variáveis, portanto, não podemos decidir sobre nenhuma variável em relação à dominância da análise.

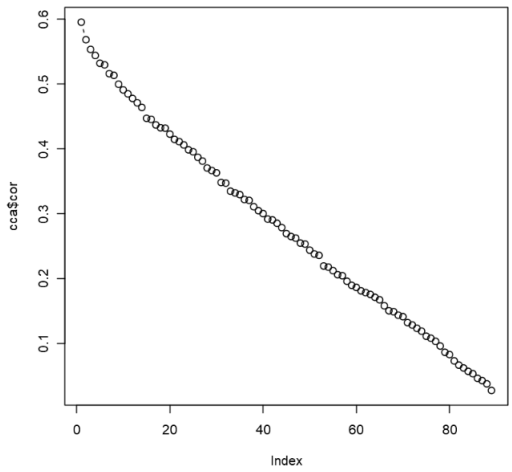

Figura 6: Gráfico dos coeficientes das correlações canônicas.

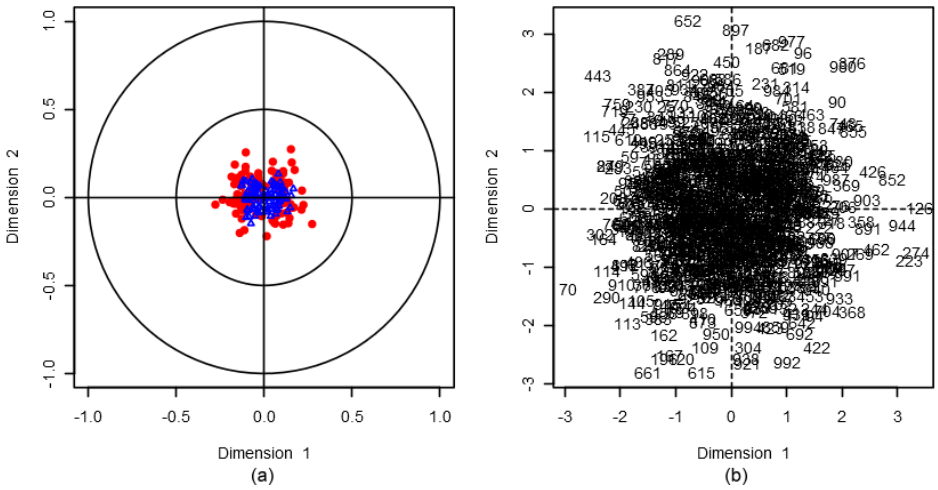

Figura 7: (a) Coeficientes das cargas para X (azul) e Y (vermelho). (b) Representação da dispersão das observações segundo a integração dos bancos de dados.

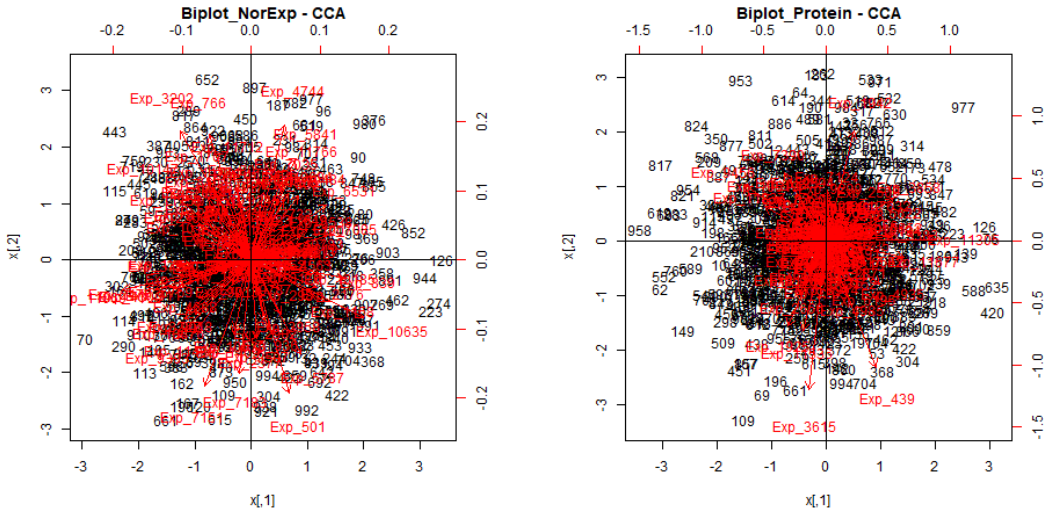

Figura 8: (a) Biplot: representação simultânea dos escores de X e Y vs cargas de Y. (b) Biplot: representação simultânea dos escores de X e Y vs cargas de X.

3.3 Discussão

Na análise do banco de dados "Protein", os dois primeiros componentes principais obtidos pela PCA explicaram apenas 15% da variância total dos dados, o que não foi suficiente para representar bem os dados em duas dimensões. Já na análise discriminante linear de Fisher, a classificação de grupos pelo método empírico apresentou um bom desempenho, alcançando uma acurácia de 81.4%, superior ao resultado obtido pelo método de validação cruzada, que obteve uma acurácia de 75.8%. Na análise de agrupamentos, o método de Ligação Completa alcançou uma acurácia de 47.84% na classificação de grupos, superior ao resultado obtido pelo método de K-Médias realizado através do algoritmo de Hartigan-Wong, que obteve uma acurácia de 45.94%. Portanto, a LDA obteve uma acurácia muito maior do que a análise de agrupamentos (tanto para o método hierárquico quanto para o não-hierárquico) na tarefa de segregação de grupos.

Na integração dos bancos de dados "Protein"e "NorExp", os dois primeiros componentes de PLS obtidos explicaram apenas 13% da variância total dos dados e constatamos pelo gráfico de correlações que a esmagadora maioria das variáveis estão pobremente correlacioandas, portanto não estão bem representadas por esses componentes PLS. Ademais, o PLS não foi um bom discriminador de grupos (escores não discriminaram os grupos). Já no estudo da CCA, obtivemos um coeficiente de correlação não muito expressivo entre as variáveis canônicas (0.5956). Ainda, obtivemos baixas correlações para a esmagadora maioria das variáveis.

4 Conclusões

Neste trabalho, mostramos que a PCA aplicada no banco de dados "Protein" não produziu uma redução de dimensionalidade satisfatória em duas dimensões, no entando a LDA conseguiu uma alta acurácia na discriminação de grupos, em detrimento da análise de agrupamentos hierárquico

e não-hierárquico que obtiveram discriminação insatisfatória. A integração dos bancos "Protein"e "NorExp", tanto pela técnica de PLS quanto pela CCA, revelou que os dois bancos de dados estão pobremente correlacionados.

Disponibilidade dos dados e códigos

Os conjuntos de dados simulados, conjuntamente com os códigos da análise, utilizados para a confecção deste artigo estão disponíveis no seguinte repositório do Github: https://github.com/heitorbaldo/Multivariate-Analysis.

Referências

[1] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons,, 2003.

[2] R. H. Chung and C. Y. Kang, *A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification.* Giga Science 8(5): 1–12, 2019.

[3] B. Everitt and T. Hothorn, *An Introduction to Applied Multivariate Analysis with R.* Springer, 2011.

[4] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis.* Wiley, 2011.

[5] R. A. Irizarry and M. I. Love, *Data Analysis for Life Sciences.* Leanpub, 2015.

[6] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis.* Pearson Prentice Hall, 2007.

[7] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis.* Academic Press, 1979.