Métodos Fatoriais de Análise de Dados e *Big Data*

Adelaide Figueiredo[1], *adelaide@fep.up.pt* e Fernanda Otília Figueiredo[2], *otília@fep.up.pt*

[1]*Faculdade de Economia da Universidade do Porto, LIAAD-INESC TEC Porto e CEAUL*
[2]*Faculdade de Economia da Universidade do Porto e CEAUL*

Introdução

Na literatura existem muitos métodos de redução de dimensão e de visualização de dados multidimensionais (*high dimensional data*). Estes métodos são importantes para transformar um grande número de variáveis correlacionadas num número menor de novas variáveis não correlacionadas entre si, e para visualizar a estrutura dos dados em espaços de dimensão menor. Em Análise de Dados, os métodos fatoriais de redução de dimensão mais populares são a Análise em Componentes Principais (ACP), a Análise das Correspondências Simples (ACS), a Análise das Correspondências Múltiplas (ACM) e a Análise Discriminante Linear (ADL). Estes métodos de estatística multidimensional têm muitas aplicações nas mais variadas áreas, como por exemplo, em saúde, engenharia, genética e ambiente, entre outras. A Análise em Componentes Principais e a Análise Discriminante Linear são também métodos muito usados nas áreas de data mining, *machine learning* e bioinformática.

Estes métodos de Análise de Dados aplicam-se a conjuntos de dados pequenos ou moderados, mas são difíceis de aplicar a *Big Data* devido a problemas de memória e armazenamento. Como se sabe hoje em dia, deparamo-nos com grandes volumes de dados, pelo que se torna necessário ver a adequabilidade dos métodos de redução de dimensão tradicionais, atrás referidos, para este tipo de dados (*big high dimensional data*). Em geral, considera-se um quadro de dados com n observações e p variáveis/caraterísticas (dimensões). Segundo Seng and Ang (2017) este quadro pode ser classificado em duas categorias: quadro com elevado número de observações (*Big Sample Data Set*) e quadro com elevado número de variáveis (*Big Feature Data Set*). Na primeira categoria ($n>>p$), a dimensão dos dados não é elevada, e à medida que o volume dos dados aumenta, o número de observações aumenta, mantendo-se o número de variáveis. Na segunda categoria ($p>>n$), o número de variáveis p é elevado, e pode ainda aumentar, tal como n pode aumentar com o crescimento do volume dos dados.

O enfoque deste trabalho é nos métodos fatoriais de redução de dimensão de Análise de Dados. Iremos referir, nas secções 2, 3 e 4, novas abordagens da Análise em Componentes Principais, da Análise das Correspondências e da Análise Discriminante Linear, respetivamente, desenvolvidas para enfrentar os desafios despoletados pela era do *Big Data*.

Análise em Componentes Principais e *Big Data*

A Análise em Componentes Principais (ACP) é um método de redução de dimensão clássico introduzido por Pearson (1901). Trata-se de um método que permite explicar a variabilidade subjacente a um conjunto de dados através de um número menor de novas variáveis não correlacionadas entre si. Estas novas variáveis, designadas por componentes principais, são combinações lineares das variáveis iniciais. O número de componentes principais que é possível determinar neste método é igual ao número de variáveis inicial, mas interessa-nos reter um número de componentes principais muito menor que o número inicial de variáveis, e que expliquem uma boa variabilidade dos dados. As representações gráficas dos indivíduos e das variáveis nos primeiros planos principais ajudam-nos a identificar a estrutura dos dados. Os resultados obtidos na Análise em Componentes Principais são em geral fáceis de interpretar, pelo que muitas vezes este método é utilizado antes de uma Análise Classificatória, de uma Regressão Linear, e de muitos outros métodos.

Na Análise em Componentes Principais começa-se por efetuar um pré-processamento da matriz dos dados, isto é, centram-se os dados de modo a que as variáveis tenham média zero; em seguida, diagonaliza-se a matriz de covariâncias para determinar as componentes principais. Em geral, as variáveis estão expressas em unidades de medida diferentes ou têm ordens de grandeza diferentes, pelo que é necessário ainda reduzir as variáveis de modo a que fiquem com desvio-padrão unitário, obtendo-se assim dados estandardizados; neste caso, diagonaliza-se em seguida a matriz de correlações entre as variáveis.

Zhang and Yang (2016) referem que as dificuldades em aplicar a Análise em Componentes Principais a *Big Data* estão relacionadas com problemas de memória e armazenamento, e propõem métodos e algoritmos para ultrapassar essas dificuldades, como iremos mencionar sucintamente em seguida. Em geral é impossível guardar a matriz de dados na memória de um computador e a estandardização das variáveis é um problema em *Big Data* porque é difícil guardar os resultados, ou na memória ou no disco duro de um computador. Ainda devido à quantidade massiva de dados que vão surgindo todos os dias, é frequente ser necessária a atualização de dados e a combinação de conjuntos de dados com os anteriores para voltarem a ser analisados. Numa abordagem clássica, temos de considerar o conjunto inteiro dos dados, o que é ineficiente num contexto de *Big Data*.

Assim, se um conjunto de dados grande pode ser guardado no disco duro de um computador, Zhang and Yang (2016) propõem um método para a Análise em Componentes Principais baseado num único processador, assumindo neste método que não pode haver estandardização. Frequentemente, não se está em condições de aplicar esta solução devido à enorme quantidade de dados que surgem diariamente, pelo que estes autores sugerem, em alternativa, recorrer a computação paralela na Análise em Componentes Principais.

Devido às dificuldades computacionais na aplicação da Análise em Componentes Principais a grandes conjuntos de dados, Halko *et al.* (2011) e Witten and Candes (2013) entre outros, aplicaram projeções de matrizes aleatórias. Estes últimos autores aproximam uma matriz de elevada dimensão pelo produto de duas matrizes de menor dimensão as quais podem ser tratadas de forma menos complicada.

Para lidar com dados não lineares, Schölkopf *et al.* (1998) propuseram a Análise em Componentes Principais Kernel (*Kernel PCA*). A matriz de covariâncias na Análise em Componentes Principais é substituída por uma matriz baseada numa função kernel.

Outros métodos de redução de dimensão têm sido propostos para visualizar *high dimensional data,* como por exemplo: o método de redução de dimensão não linear, Kernel Entropy Components Analysis (KECA), proposto por Jenssen (2010), no qual se maximiza a entropia quadrática de Renyi; técnicas de redes neuronais (*Deep neural network*) tais como *Deep Belief Network* (DBN), referido em Noulas and Krse (2008), ou *Staked Auto-encoders* (SAE), apresentado em Schmidhuber (2015).

Tsai (2011) usou a Análise em Componentes Principais para redução de dimensão, de forma a efetuar a visualização de outliers. Najim and Lim (2014) usaram a Análise em Componentes Principais como um método de redução de dimensão para avaliar a qualidade de visualização.

Zhan and El Ghaoui (2011) mostram que, na prática, a Análise em Componentes Principais esparsa (*sparse PCA*) pode ser mais simples que a ACP, e pode ser aplicada a conjuntos de dados muito grandes, como por exemplo, a dados textuais envolvendo milhões de documentos e com centenas de milhares de caraterísticas.

Análise das Correspondências e *Big Data*

A Análise das Correspondências permite estudar as relações entre duas variáveis qualitativas (Análise das Correspondências Simples) ou entre mais do que duas variáveis qualitativas (Análise das Correspondências Múltiplas).

A Análise das Correspondências de um número infinito de linhas ou observações por 1000 atributos foi abordada por Benzécri (1982, 1997). Murtagh (2015a) aplica a Análise das Correspondências a *Big Data*, isto é, a 30 milhões de palavras, apresentando os enormes outputs da análise de forma "inteligente". Murtagh (2015b) descreve propriedades úteis de espaços de dados com elevada dimensionalidade, as quais podem ter interesse na análise de *Big Data*.

Uma nova abordagem da Análise das Correspondências Múltiplas baseada em redes neuronais foi proposta por Tian and Chen (2017) para deteção de falhas em sistemas de gestão de informação.

Análise Discriminante Linear e *Big Data*

A Análise Discriminante Linear (ADL) é um método de redução de dimensão que tem por objetivo determinar as combinações lineares das variáveis iniciais que melhor discriminam os grupos de observações definidos à partida.

Este método também tem por objetivo classificar um novo indivíduo numa de várias classes com base nos valores observados das variáveis para esse indivíduo. Como a distribuição de probabilidade das variáveis é geralmente desconhecida, a regra de classificação é construída usando uma amostra treino.

Muitas vezes em problemas de classificação aplicam-se os dois métodos, ACP e ADL: começa-se por aplicar a ACP para reduzir a dimensionalidade dos dados, e em seguida, aplica-se a ADL. Contudo com *Big Feature Data Sets* pode não ser adequado aplicar a ACP antes da ADL, porque se pode perder poder discriminatório na ADL. A Análise Discriminante Linear para *Big Data*, em geral, ou especificamente para *Big Feature Data Sets*, não tem sido completamente explorada. No entanto, Seng and Ang (2017) propuseram uma abordagem designada por *Split and Combine Linear Discriminant Analysis* (SC-LDA) para *Big Feature Data Analytics*. Contrariamente às abordagens da Análise Discriminante Linear e suas extensões, em que o objetivo é essencialmente melhorar a velocidade e eficiência dos cálculos envolvidos no método, a abordagem SC-LDA tem por objetivo não só reduzir o custo de computação, como também dividir o problema da Análise Discriminante Linear em dois sub-problemas de dimensão menor, resolver os sub-problemas separadamente com um algoritmo base, e combinar os resultados dos dois sub-problemas para obter a solução final. Tal como a ACP, a Análise Discriminante Linear clássica requer a diagonalização de uma matriz que é muito dispendiosa computacionalmente. A abordagem SC-LDA substitui a diagonalização dessa matriz pela diagonalização de sub-matrizes mais pequenas que podem ser efetuadas em paralelo, e depois combina os resultados intermédios para obter os resultados da diagonalização da matriz global.

Existem outras abordagens recentes da Análise Discriminante Linear, tais como algoritmos de ADL de aprendizagem incremental (*Incremental Learning LDA algorithms*) e ADL para conjuntos de dados com poucas observações (*LDA for undersampled data sets*). Os algoritmos de ADL de aprendizagem incremental têm a vantagem de lidar com os novos dados que vão surgindo, i.e, são facilmente aplicados a *data streams*, e a aprendizagem de todos os dados desde o início não é requerida. Isto não exige elevada complexidade computacional e o sistema não necessita de muita capacidade de memória para armazenar os dados, quer aprendidos anteriormente, quer apresentados de novo. Algoritmos de ADL deste tipo foram propostos, por exemplo, por Uray *et al.* (2007), Kim *et al. (*2011) e Ghassabeh and Moghaddam (2013).

Os algoritmos de ADL para *undersampled data sets* pretendem resolver um problema bem conhecido na ADL clássica, designado por problema de singularidade, o qual ocorre quando o número de variáveis *p* na matriz de dados é elevado comparado com o número *n* de observações (*Big feature data set*). Outras abordagens para contornar o problema da singularidade têm sido propostas, as quais consistem em usar diferentes variantes da função objetivo da ADL clássica, como por exemplo, em Chu *et al.* (2011). Shao *et al.* (2011) propõem uma Análise Discriminante Linear esparsa (*sparse LDA*) para o caso em que o número de variáveis usada para a classificação é muito maior que a dimensão da amostra, uma vez que neste caso a ADL pode ter uma *performance* não adequada. Qiao *et al.* (2009) também desenvolvem um procedimento de *sparse LDA* eficaz para o caso de *high dimensional data* e amostras de dimensão pequena.

Ye and Wang (2006) apresentam um novo algoritmo para a Análise Discriminante Regularizada (ADR) no caso de *high dimensional data*. É de relembrar que a ADR foi proposta por Friedman (1989) como um compromisso entre a Análise Discriminante linear e quadrática, e tem mostrado ser flexível para lidar com várias classes de distribuições.

Referências

Benzécri, J. P. (1982). L'approximation stochastique en analyse des correspondances, *Les Cahiers de l'Analyse des Données*, **7**(4), 387-394.

Benzécri, J. P. (1997). Approximation stochastique, réseaux de neurones et analyse des données, *Les Cahiers de l' Analyse des Données*, **22**(2), 211-220.

Chu, D., Goh, S. T. and Hung, Y. S. (2011). Characterization of all solutions for undersampled uncorrelated linear discriminant analysis problems, *SIAM Journal on Matrix Analysis and Applications*, **32**(3), 820-844.

Friedman, J. H. (1989). Regularized discriminant analysis, *Journal of the American Statistical Association*, **84**(405), 165-175.

Ghassabeh, Y. A. and Moghaddam, H. A. (2013). Adaptive linear discriminant analysis for online feature extraction, *Machine Vision and Applications*, **24**(4), 777-794.

Halko, N., Martinsson, P. G. and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, *SIAM Review*, **53**(2), 217-288.

Jenssen, R. (2010). Kernel entropy component analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **32**(5), 847-860.

Qiao, Z., Zhou, L. and Huang, J. Z. (2009). Sparse Linear Discriminant Analysis with Applications to High Dimensional Low Sample Size Data, *IAENG International Journal of Applied Mathematics*, **39**(1), 48-60.

Murtagh, F. (2015a). Correspondence Factor Analysis of Big Data Sets: A Case Study of 30 Million Words; and Contrasting Analytics using Apache Solr and Correspondence Analysis in R, *Technical Report, Goldsmiths University of London*, 1-38.

Murtagh, F. (2015b). Big Data Scaling through Metric Mapping: Exploiting the Remarkable Simplicity of Very High Dimensional Spaces using Correspondence Analysis. *Technical Report, Goldsmiths University of London*, 1-13.

Najim, S. A. and Lim, I. S. (2014). Trustworthy dimension reduction for visualization different data sets, *Information Sciences*, 278, 206-220.

Noulas, A. K. and Krse, B. J. A. (2008). Deep Belief Networks for dimension reduction. In Proceedings of *Belgian-Dutch Conference on Artificial Intelligence*, Netherland, 20, 185-191.

Kim, T., Wong, S., Stenger, B., Kittler, J. and Cipolla, R. (2011). Incremental linear discriminant analysis using sufficient spanning sets and its applications, *International Journal of Computer Vision*, **9**(2), 216-232.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space, *Philosophical Magazine*, Series 6, **2**(11), 559-572.

Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview, *Neural Networks*, 61, 85-117.

Schölkopf, B., Smola, A. and Müller, K.-R. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem, *Neural Computation*, July 1, **10**(5), 1299-1319.

Seng, J. K. P. and Ang, K. L.-M. (2017). Big Feature Data Analytics: Split and Combine Linear Discriminant Analysis (SC-LDA) for Integration Towards Decision Making Analytics, *IEEE Access*, 5, 14056-14065.

Shao, J., Wang, Y., Deng, X. and Wang, S. (2011). Sparse Linear Discriminant Analysis by thresholding for high dimensional data*, The Annals of Statistics*, 39, 2, 1241-1265.

Tian, H. and Chen, S.-C. (2017). MCA-NN: Multiple Correspondence Analysis based Neural Network for Disaster Information Detection, *IEEE Third International Conference on Multimedia Big Data*, 268-275.

Tsai, F. S. (2011). Dimensionality reduction techniques for blog visualization, *Expert Systems with Applications*, **38**(3), 2766-2773.

Uray, M., Skocaj, D., Roth, P. M. and Bischof, A. L. H. (2007). Incremental LDA learning by combining reconstructive and discriminative approaches, in *Proceedings of British Machine Vision Conference (BMVC)*, 272-281.

Witten, R. and Candes, E. (2013). Randomized algorithms for low–rank matrix factorizations: Sharp performance bounds, *Algorithmica*, **72**(1), 264-281.

Ye, J. and Wang, T. (2006). Regularized Discriminant Analysis for High Dimensional Low Sample Size Data, *International Conference* on *Knowledge Discovery and Data Mining* (*KDD*), Philadelphia, PA, USA, 454-463.

Zhang, Y. and El Ghaoui, L. (2011). Large-Scale Sparse Principal Component Analysis with Application to Text Data, in *Proceedings Advances in Neural Information Processing Systems 24 (NIPS)*, 1-8.

Zhang, T. and Yang, B. (2016). Big Data Dimension Reducing using PCA, *IEEE International conference on Smart Cloud*, 152-157.