PEL208 — Relatório Atividade 2
Implementação da Análise de Componentes Principais

Miller Horvath
Mestrando em Engenharia Elétrica (Processamento de Sinais e Imagens)
Centro Universitário FEI, São Bernardo do Campo, SP, Brasil

23 de outubro de 2018

1 Introdução

Este trabalho apresenta o relatório do desenvolvimento da segunda atividade avaliativa referente à disciplina PEL208, intitulada Tópicos Especiais em Aprendizagem, apresentada pelo Prof. Dr. Reinaldo Augusto da Costa Bianchi.

O objetivo desta atividade é implementar a Análise de Componentes Principais (PCA), conforme abordado em sala de aula. Para isso, a linguagem de programação *C++* foi adotada. A atividade permite a utilização de bibliotecas que proveem operações de álgebra linear, especialmente para calcular os autovalores e autovetores de uma matriz.

2 Conceitos Fundamentais

2.1 Análise de Componentes Principais (PCA)

A técnica de PCA, primeiramente formulada em (PEARSON, 1901) e posteriormente desenvolvida para a forma como conhecemos hoje em (HOTELLING, 1933), aplica uma transformação ortogonal em um conjunto de dados, no qual as variáveis podem ou não ser correlacionadas, convertendo-o em um novo conjunto de dados representado por variáveis linearmente independentes, chamadas de componentes principais (CP).

Desta forma, cada CP é composta por uma combinação linear das variáveis do conjunto de dados original (RINGNÉR, 2008). A primeira CP é calculada de modo a representar a maior variância possível dos dados originais, a segunda CP deve ser ortogonal a primeira, e as demais CPs são calculadas de maneira similar (ABDI; WILLIAMS, 2010). O método tradicional para determinar as CPs calcula os autovetores (a) e autovalores (λ) da matriz de covariância (S) do conjunto de dados, através da Equação 1 presente em (JOLLIFFE; CADIMA, 2016); sendo que a quantidade de CP para um conjunto

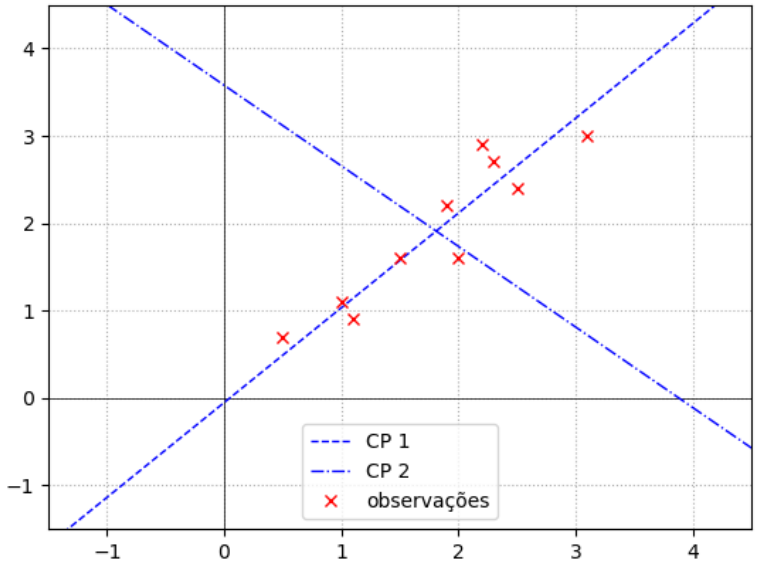

Figura 1 – Representação gráfica do cálculo das componentes principais de um conjunto de dados bi-dimensional.

de dados é sempre igual ao número de variáveis contidas no mesmo e apenas matrizes quadradas possuem autovetores, mas nem toda a matriz quadrada possui autovetores. A Figura 1 ilustra o cálculo do PCA para um conjunto de dados bi-dimensional.

$$Sa - \lambda a = 0 \iff Sa = \lambda a \tag{1}$$

Uma das principais aplicações do PCA é a redução de dimensionalidade, onde o conjunto de dados passa a ser representado pelas CPs mais importantes, ou seja, aquelas que representam maior variabilidade dos dados, dado que a importância de uma CP é determinada pelo seu autovalor correspondente. Quando todas as CPs são utilizadas, é possível reconstruir os dados originais perfeitamente. Por outro lado, ainda é possível reconstruir os dados utilizando um número limitado de CPs, porém, quanto menor é a quantidade de CPs utilizadas, maior é perda de informações. A Figura 2 mostra o conjunto de dados utilizado na Figura 1 reconstruídos utilizando apenas a CP de maior importância.

Ademais, o método de decomposição em valores singulares (SVD) é bastante relacionado com o PCA, podendo-se utilizar o SVD como uma forma generalizada para solução do PCA (SHLENS, 2014).

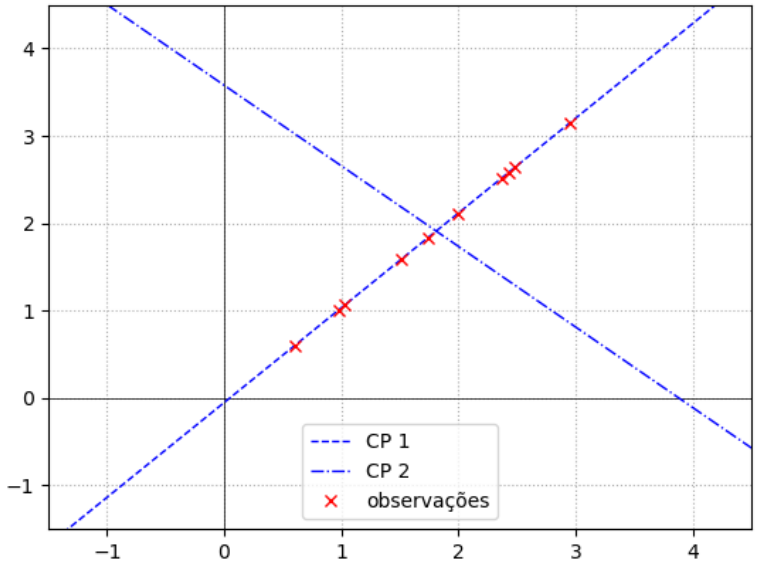

Figura 2 – Representação gráfica da reconstrução dos dados utilizando apenas uma componente principal.

3 Trabalhos Relacionados

O PCA possuí diversas aplicações, tais como redução de dimensionalidade, seleção de variáveis, detecção de *outliers*, classificação, predição, entre outras (WOLD; ESBENSEN; GELADI, 1987). Nesta seção, são apresentados alguns trabalhos acadêmicos recentes que utilizam PCA.

Em (ZHAO; SHKOLNISKY; SINGER, 2016), é proposto o *fast Fourier-Bessel steerable PCA (FFBsPCA)*, que otimiza o método proposto em (ZHAO; SINGER, 2013) através do cálculo mais rápido e preciso dos coeficientes de expansão de *Fourier-Bessel*, diminuindo a complexidade computacional do método de $O(nL^4)$ para $O(nL^3)$.

Em (JOLLIFFE; CADIMA, 2016), foi realizada uma revisão em PCA, onde algumas variações do PCA são apresentadas. Essas variações podem visar adaptar o PCA para objetivos diferentes ou buscar analisar tipos de dados específicos (JOLLIFFE; CADIMA, 2016).

Em (METSALU; VILO, 2015), é proposta uma plataforma web que oferece uma interface para que pesquisadores com poucas habilidades de programação possam aplicar o PCA em suas bases de dados.

Em (QURESHI et al., 2017), PCA é aplicado em bases de dados médicos com o objetivo de explorar as componentes mais relevantes para o aumento de risco de doenças cardíacas isquêmicas.

4 Metodologia

A implementação da atividade foi desenvolvida na linguagem *C++*, utilizando o software Visual Studio 2017. Para apoiar os cálculos algébricos, foram utilizados os seguintes recursos da biblioteca *Eigen* (GUENNEBAUD; JACOB et al., 2010):

- Classe *MatrixXd*, sendo uma estrutura de dados para matrizes multidimensionais compostas por valores do tipo *double*;
- Classe *VectorXd*, sendo uma estrutura de dados para vetores multidimensionais compostos por valores do tipo *double*;
- Sobrecarga de operadores para soma, subtração; divisão e multiplicação de vetores e matrizes;
- Métodos de transposição e inversão de matrizes;
- Classe *EigenSolver*, para calcular os autovetores e autovalores de uma matrix.

Foi desenvolvida uma classe chamada PCA, que recebe a matriz de dados A como parâmetro de sua função construtora para que a análise de componentes principais seja aplicada em A. Após o cálculo do PCA, a classe armazena os seguintes atributos privados:

- Dados da matriz A normalizados com média 0 (zero);
- Os autovalores de A (ordenados de forma decrescente)
- Os autovetores de A (ordenados de forma decrescente em relação aos seus respectivos autovalores)
- A variância explicada por cada componente principal de A (que são os autovalores normalizados para representar a variância proporcional explicada por cada componente principal)
- As médias originais das variáveis em A
- A matriz de covariância de A.

Além disso, a classe também oferece as seguintes funções:

- As funções *getters* para acessar os atributos privados;
- Construtor de cópia, que recebe como parâmetro um objeto instanciado da classe PCA;
- Sobrecarga do operador de atribuição "=";

- Função *transform* — recebe como parâmetro um valor inteiro c e aplica uma redução de dimensionalidade em A, retornando uma nova matriz com c variáveis;
- Função *rebuild* — recebe como parâmetro uma matriz B, resultante da função *transform*, e retorna uma matriz com a quantidade de variáveis originais de A resultante reconstruída a partir dos dados em B.

5 Experimentos

Para testar e analisar a implementação desenvolvida, foram utilizadas 3 bases de dados, que serão descritas nas subseções seguintes. Em cada uma das bases de dados, foi utilizado o PCA, descrito na Seção 2.1, para encontrar as suas componentes principais. Ademais, a primeira componente principal é comparada com a curva resultante da regressão através do método dos mínimos quadrados.

Para a base *Hald Dataset* 5.1.4, será aplicado o PCA para calcular os seus autovetores e determinar a quantidade de CPs que são necessárias para representar ao menos 98% da variância dos dados.

5.1 Bases de Dados

5.1.1 Alps Water

Possuí informações sobre o ponto de ebulição da água sob efeito de diferentes pressões atmosféricas. Esta base de dados possuí 17 observações e 2 características, que são:

- Temperatura — temperatura do ponto de ebulição da água medida em graus Fahrenheit (F); e
- Pressão — pressão atmosférica medida em polegadas de mercúrio (”Hg).

Sendo que, nesta atividade, a pressão foi definida como a variável objetivo desta base de dados

5.1.2 Livros x Notas

Possuí informações sobre o desempenho de alunos em uma aula de estatística. Esta base de dados possuí 40 observações e 3 características, que são:

- Livros — quantidade de livros lidos pelos alunos;
- Assiduidade — quantidade de aulas em que os alunos estavam presentes; e
- Nota — nota final dos alunos no curso.

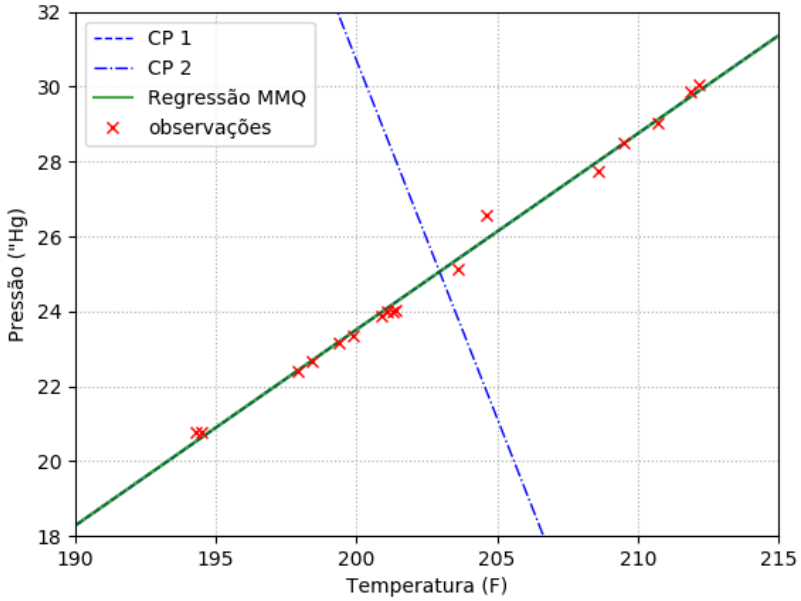

Figura 3 – Ilustração que compara as componentes principais com a curva de regressão linear (método dos mínimos quadrados) da base de dados *Alps Water* (Seção 5.1.1), conforme descrito na Seção 5

Sendo que, nesta atividade, a nota foi definida como a variável objetivo desta base de dados

5.1.3 US Cesus Dataset

Possuí informações sobre o registro de contagem populacional nos Estados Unidos a cada 10 anos. Esta base de dados possuí 11 observações e 2 características, que são:

- Ano — ano em que a contagem foi realizada; e
- População — numero de habitantes registrados.

Sendo que, nesta atividade, a população foi definida como a variável objetivo desta base de dados

5.1.4 Hald Dataset

Esta base de dados possuí 13 observações, 4 características, e 1 variável objetivo.

6 Resultados

Nesta seção, são apresentados os gráficos obtidos através do desenvolvimento experimental descrito na Seção 5.

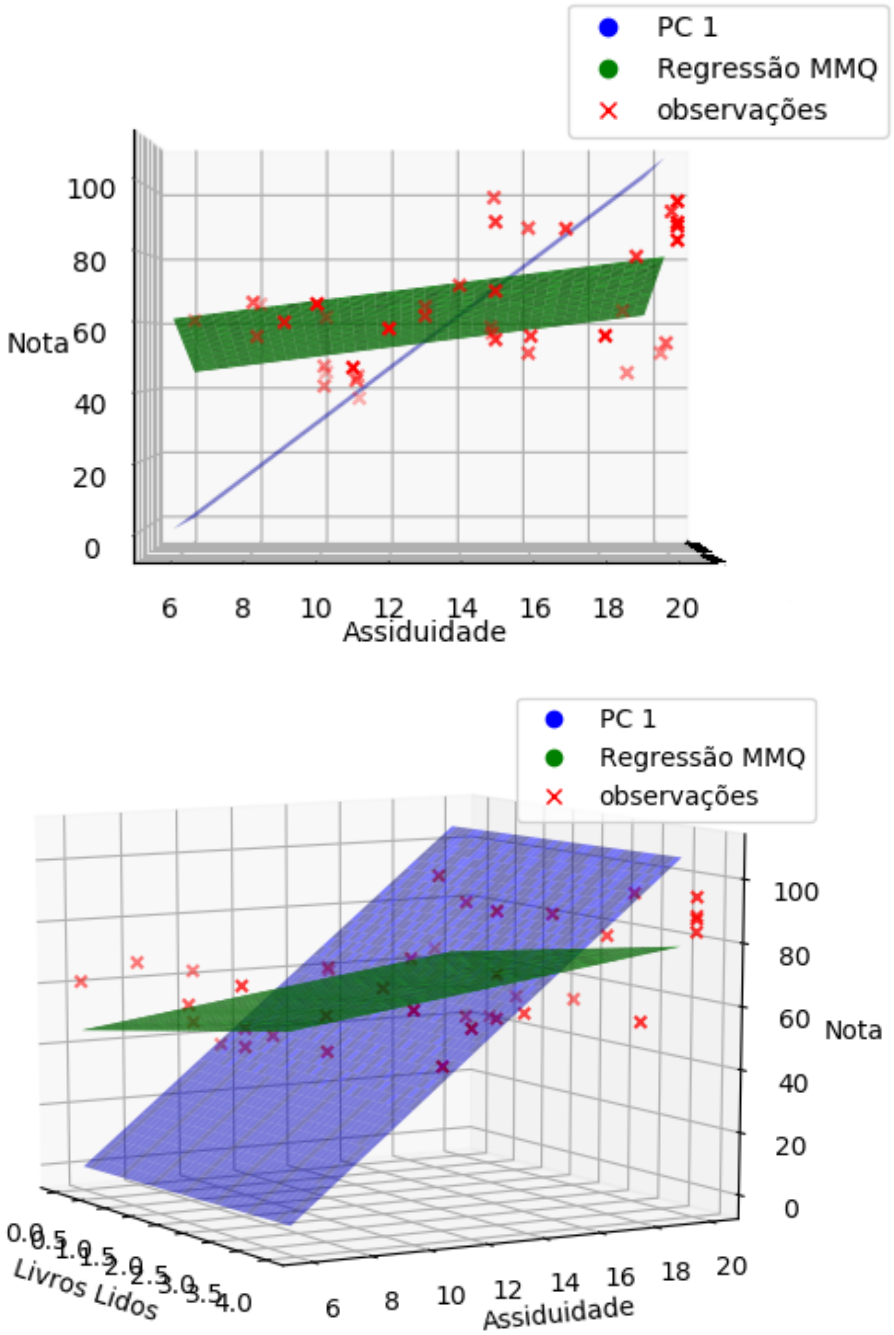

Figura 4 – Ilustração que compara as componentes principais com a curva de regressão linear (método dos mínimos quadrados) da base de dados *Livros x Notas* (Seção 5.1.2), conforme descrito na Seção 5

6.1 Alps Water

A Figura 3 mostra que a primeira componente principal e a reta resultante da regressão linear são quase idênticas. Isso ocorreu nesta base devido à alta dependência linear entre as variáveis.

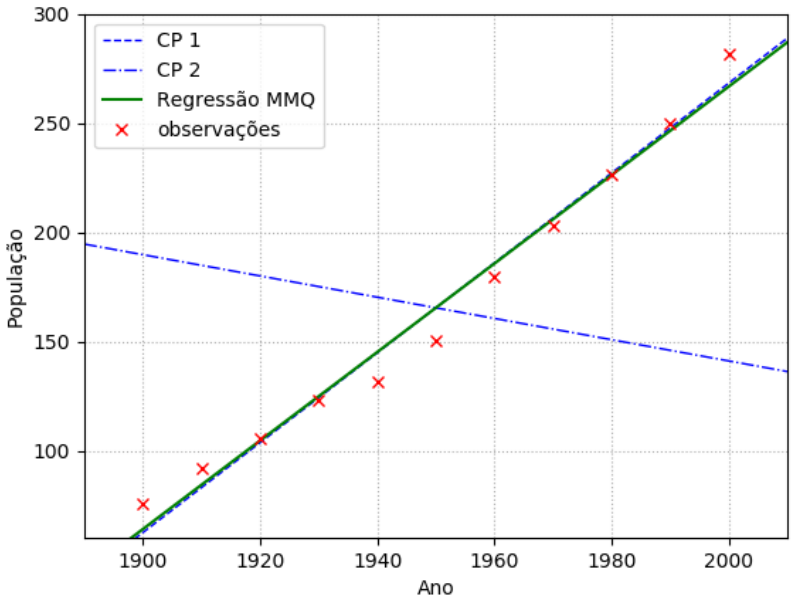

Figura 5 – Ilustração que compara as componentes principais com a curva de regressão linear (método dos mínimos quadrados) da base de dados *US Census Dataset* (Seção 5.1.3), conforme descrito na Seção 5

6.2 Livros x Notas

A Figura 4 mostra que a primeira componente principal e o plano resultante da regressão linear são bem diferentes. Neste caso, isso ocorre devido a diferença característica de cada um dos modelos. A primeira CP mostra o plano onde ocorre a maior variância dos dados. Já a regressão linear, apresenta o plano que minimiza os resíduos entre as observações e o próprio plano.

6.3 US Census Dataset

Similarmente ao demonstrado na Figura 3, a Figura 5 mostra que a primeira componente principal e a reta resultante da regressão linear também são muito parecidas.

7 Hald Dataset

Conforme descrito na Seção 5, foi aplicado o PCA na base de dados *Hald Dataset*, descrita na Seção 5.1.4, para encontrar os seus autovetor e determinar o número de componentes principais necessário para representar 98% da variância dos dados. Dessa forma, a Tabela 1 apresenta as componentes principais (autovetores) resultantes do PCA.

A Tabela 2 apresenta os autovalores equivalentes aos autovetores apresentados na Tabela 1. A Tabela 3 apresenta a porcentagem da variância total dos dados que são

CP 1	CP 2	CP 3	CP 4
-0.0678	-0.646018	0.567315	0.50618
-0.678516	-0.0199933	-0.543969	0.493268
0.0290208	0.75531	0.403553	0.515567
0.730874	-0.10848	-0.468398	0.484416

Tabela 1 – Autovetores (componentes principais) obtidos através da aplicação do PCA na base de dados *Hald Dataset* (Seção 5.1.4).

Autovalor 1	517.797
Autovalor 2	67.4964
Autovalor 3	12.4054
Autovalor 4	0.237153

Tabela 2 – Autovalores equivalentes as CPs apresentadas na Tabela 1.

Componente Principal	Variância Representada
1	0.865974
2	0.112882
3	0.0207471
4	0.00039662

Tabela 3 – Porcentagem da variância representada por cada CP da base de dados *Hald Dataset* (Seção 5.1.4).

representadas por cada uma das CPs. Desta forma, 1 CP representa aproximadamente 87% da variância, 2 CPs representam aproximadamente 98% da variância, 3 CPs são suficientes para representar quase 100% da variância e a última CP possuí uma representatividade irrisória da variância. Portanto, 2 CPs são suficientes para representar 98% da variância dos dados nesta base.

8 Conclusão

Este trabalho visa relatar a implementação da análise de componentes principais, utilizando *C++* como linguagem de programação, como tarefa do curso PEL208 do programa de pós-graduação em engenharia elétrica do Centro Universitário FEI. O fato dos resultados, apresentados na Seção 6, serem condizentes com os apresentados em aula sugere que a implementação foi adequada.

O PCA é uma técnica que aplica uma transformação ortogonal em um conjunto de dados de modo a identificar as componentes principais que mais explicam a variância destes dados. Esta representatividade é frequentemente utilizada para aplicar uma redução de dimensionalidade no conjunto de dados minimizando a perda de informações.

A primeira componente principal foi comparada com a curva resultante da regressão linear dos dados. Esta comparação destaca as diferenças dos dois métodos estatísticos que,

apesar de resultarem em curvas semelhantes quando há uma grande dependência linear entre as variáveis, analisam características distintas do conjunto de dados.

Referências

ABDI, H.; WILLIAMS, L. J. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, Wiley Online Library, v. 2, n. 4, p. 433–459, 2010. Citado na página 1.

GUENNEBAUD, G.; JACOB, B. et al. *Eigen v3*. 2010. Http://eigen.tuxfamily.org. Citado na página 4.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, Warwick & York, v. 24, n. 6, p. 417, 1933. Citado na página 1.

JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc. A*, The Royal Society, v. 374, n. 2065, p. 20150202, 2016. Citado 2 vezes nas páginas 1 e 3.

METSALU, T.; VILO, J. Clustvis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. *Nucleic acids research*, Oxford University Press, v. 43, n. W1, p. W566–W570, 2015. Citado na página 3.

PEARSON, K. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Taylor & Francis, v. 2, n. 11, p. 559–572, 1901. Citado na página 1.

QURESHI, N. A. et al. Application of principal component analysis (pca) to medical data. *Indian Journal of Science and Technology*, v. 10, n. 20, 2017. Citado na página 3.

RINGNÉR, M. What is principal component analysis? *Nature biotechnology*, Nature Publishing Group, v. 26, n. 3, p. 303, 2008. Citado na página 1.

SHLENS, J. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014. Citado na página 2.

WOLD, S.; ESBENSEN, K.; GELADI, P. Principal component analysis. *Chemometrics and intelligent laboratory systems*, Elsevier, v. 2, n. 1-3, p. 37–52, 1987. Citado na página 3.

ZHAO, Z.; SHKOLNISKY, Y.; SINGER, A. Fast steerable principal component analysis. *IEEE transactions on computational imaging*, IEEE, v. 2, n. 1, p. 1–12, 2016. Citado na página 3.

ZHAO, Z.; SINGER, A. Fourier–bessel rotational invariant eigenimages. *JOSA A*, Optical Society of America, v. 30, n. 5, p. 871–877, 2013. Citado na página 3.