ISSN 2177-9139

XX EREMAT - Encontro Regional de Estudantes de Matemática da Região Sul
Fundação Universidade Federal do Pampa (UNIPAMPA), Bagé/RS, Brasil. 13-16 nov. 2014.

XX EREMAT SUL

Encontro Regional de Estudantes de Matemática da Região Sul

ANÁLISE DE COMPONENTES PRINCIPAIS: CONSTRUÇÃO VIA ÁLGEBRA LINEAR

Luana da Silva – luds6851@gmail.com
Universidade Federal de Santa Maria, Campus Santa Maria, 97105-900, Santa Maria, RS, Brasil
Alice de Jesus Kozakevicius - alicek@ufsm.com
Universidade Federal de Santa Maria, Campus Santa Maria, 97105-900, Santa Maria, RS, Brasil

Resumo: *A análise de componentes principais (ACP) é uma técnica estatística muito utilizada que tem aplicações em vários campos de pesquisa tais como compressão de imagens e análise de eletrocardiogramas. Além disso, ela ainda é bastante utilizada para determinar padrões em dados com dimensões muito altas. Este trabalho sintetiza os conceitos da ACP, dando enfoque na formulação matemática envolvida, como a obtenção da matriz de covariância a partir dos dados iniciais, a determinação de seus autovalores e autovetores, a construção da nova base de representação dos dados a partir da análise desses autovetores, etc. Ao final são descritas duas aplicações da análise de componentes principais.*
Palavras-chave: *Análise de Componentes Principais, Reconhecimento de Padrões, Autovalores, Autovetores.*

1. INTRODUÇÃO

Quando se tem contato com álgebra linear em um primeiro curso, raramente o enfoque são as aplicações em outras áreas que não a própria matemática, uma vez que todo o conteúdo precisa ser apresentado e definido. No entanto, a partir de conceitos básicos, em especial produto interno, dependência e independência linear, gerador de um subespaço vetorial, autovalores e autovetores, métodos importantes de classificação e reconhecimento de padrões já podem ser construídos, dentre eles o método de análise de componente principal (ACP), cujo objetivo é: a partir de um conjunto inicial de dados que possuam algum tipo de dependência, encontrar um subconjunto (de componentes principais) com a "menor redundância possível" (ou seja, sendo linearmente independente), que represente da melhor forma possível esse conjunto de dados iniciais. Naturalmente, a qualidade da representação com "a menor redundância possível" e "da melhor forma possível" dependerá do contexto no qual a aplicação original é apresentada.

Uma das áreas de aplicação que mais têm crescido por causa do desenvolvimento tecnológico é a medicina. O aumento na qualidade dos recursos computacionais impulsionou o desenvolvimento de métodos numéricos (estatísticos e matemáticos) e consequentemente, aumentou a confiabilidade das análises realizadas para reconhecimento e classificação de padrões. Por exemplo, os sinais biológicos de natureza nervosa (sinais cerebrais, sinais cardíacos...), que podem ser medidos como sinais elétricos, têm sido muito úteis na

identificação de padrões associados a patologias e seu estudo através de técnicas como ACP se torna relevante para o desenvolvimento da área.

Por ser um método muito utilizado em aplicações biomédicas e em várias outras áreas, este trabalho tem o objetivo de divulgá-lo e apresentá-lo através da álgebra linear, aproximando a linguagem considerada na formulação estatística do que é considerado no contexto de álgebra linear. Além disso, algumas aplicações são apresentadas no final.

2. MÉTODO

O método de análise de componentes principais (ACP) pode ser interpretado como sendo a obtenção de uma transformação linear ortogonal que leva os dados iniciais do problema para um novo sistema de coordenadas, identificando quais são as direções de maior relevância para representar os dados em questão.

Há alguns modos de se encontrar as componentes principais associadas a um conjunto de dados discretos, entre os quais há a transformada Karhunen-Leòve (Wikipédia, 2014) e a formulação baseada na matriz de covariância (Richardson, 2009), a qual iremos abordar neste texto nas seções que seguem.

Vamos supor que sejam obtidas *m* amostras de dados de um certo problema e que cada uma delas esteja associada a um vetor com *n* elementos. Por exemplo, as respostas de $m = 10$ pessoas para um questionário de $n = 3$ perguntas em uma pesquisa qualquer. Assim, formamos uma matriz *A* com *m* linhas e *n* colunas, armazenando esses vetores. Em álgebra linear, esse número *n* de informações (i.e., as características) distintas que representam as amostras (i.e., as pessoas) é dito a dimensão do espaço vetorial no qual iremos representar esses 3 atributos.

De um modo geral, o que queremos fazer agora é transformar linearmente, através de uma matriz P de mudança de base de dimensão *n* x *n*, cada uma das amostras contidas nas linhas da matriz *A* (que guardaremos nas colunas de A^T) em outra matriz *Y*, *n* x *m*, que conterá os dados iniciais escritos em um novo referencial,

$$Y = PA^T \tag{1}$$

A matriz *Y* conterá as amostras de *A* em um novo sistema de coordenadas, proporcionando uma nova representação dos dados. A matriz *P* pode ser vista como uma matriz de rotação que transforma *A* em *Y*, evidenciando quais direções são as preferenciais. O que vamos mostrar a seguir é como obter essa matriz *P* que realiza esta mudança de base, através da análise espectral da matriz de covariância construída a partir de *A*.

2.1 Matriz de Covariância

2.1.1 Matriz de dados

Em aplicações práticas, as amostras podem estar associadas a dados muito correlacionados, ou seja, os vetores mesmo sendo distintos podem estar muito próximos de serem linearmente dependentes (LD). Assim, uma alternativa é procurarmos descorrelacionar os dados originais encontrando direções (dentre as *n* possíveis estipuladas pelo problema) nas quais a variância é maximizada. Para isso, ao invés de *A*, vamos considerar a matriz de covariância dos dados de *A*.

Pensando na situação de termos 10 amostras relacionadas às respostas de 10 pessoas a um questionário de 3 perguntas, vamos obter a matriz *C*, *3* x *3*, de covariância de A da seguinte maneira: suponhamos que as características escolhidas sejam *idade*, *peso* (kg), *altura* (m), e

cada linha contenha as informações de cada entrevistado nesta ordem. Na equação (1) é dada a matriz *A* com estas informações.

Vamos inicialmente calcular a média de cada uma das 3 colunas. Teremos uma média das idades (*i_med* = 35,3 anos), uma média dos pesos (*p_med* = 73,6 kg) e uma média das alturas (*a_med* = 1,666 m). Como esses valores médios são todos diferentes de zero, uma maneira de se padronizar os dados é deixá-los todos com média zero em relação a cada uma das 3 direções consideradas. Para isso, é preciso subtrair a média *i_med* de cada um dos valores da primeira coluna, o peso *p_ped* de cada um dos valores da segunda coluna e a altura *a_med* dos valores da terceira coluna para cada uma das 10 pessoas, como é mostrado na equação (2) abaixo, gerando a matriz *X*:

$$A = \begin{pmatrix} 28 & 60 & 1.59 \\ 33 & 68 & 1.70 \\ 45 & 71 & 1.68 \\ 17 & 79 & 1.69 \\ 25 & 57 & 1.65 \\ 47 & 83 & 1.71 \\ 19 & 47 & 1.60 \\ 49 & 92 & 1.71 \\ 36 & 90 & 1.72 \\ 54 & 89 & 1.61 \end{pmatrix}, \; X = \begin{pmatrix} 28 - i\_med & 60 - p\_med & 1.59 - a\_med \\ 33 - i\_med & 68 - p\_med & 1.70 - a\_med \\ 45 - i\_med & 71 - p\_med & 1.68 - a\_med \\ 17 - i\_med & 79 - p\_med & 1.69 - a\_med \\ 25 - i\_med & 57 - p\_med & 1.65 - a\_med \\ 47 - i\_med & 83 - p\_med & 1.71 - a\_med \\ 19 - i\_med & 47 - p\_med & 1.60 - a\_med \\ 49 - i\_med & 92 - p\_med & 1.71 - a\_med \\ 36 - i\_med & 90 - p\_med & 1.72 - a\_med \\ 54 - i\_med & 89 - p\_med & 1.61 - a\_med \end{pmatrix} \quad (2)$$

Agora, cada uma das *n* = 3 colunas de *X* possui vetores com média zero, e portanto os dados ficaram agora todos dispersos em torno da origem. A seguir a Figura 1 apresenta um gráfico que representa os dados originais nas três dimensões do problema: idade, peso e altura.

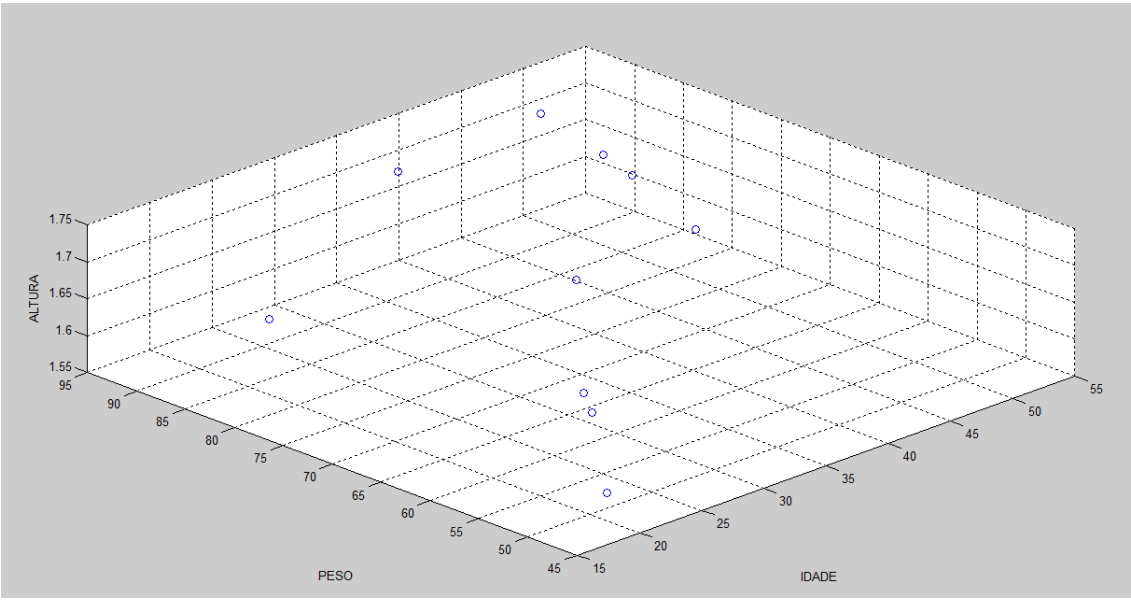

Figura 1: Gráfico da representação dos dados

2.1.2 Cálculo da matriz de covariância

A matriz de covariância *C*, de dimensão *3* x *3* descreve o quão relacionados ou não (dependentes ou independentes) são os dados das colunas da matriz *X*, ressaltando a variação

entre eles (Wikipédia, 2014). Podemos pensar na covariância como sendo uma medida estatística próxima do que é o produto interno. Por definição, a matriz de covariância para um conjunto de dados com *n* = 3, no exemplo apresentado inicialmente é:

$$C = \frac{1}{n-1} X^T X = \begin{pmatrix} < idade, idade > & < idade, peso > & < idade, altura > \\ < peso, idade > & < peso, peso > & < peso, altura > \\ < altura, idade > & < altura, peso > & < altura, altura > \end{pmatrix} \quad (3)$$

O fator $\frac{1}{n-1}$ multiplicando a matriz de produtos internos é uma definição para matriz de covariância usada quando os dados representam uma amostra não muito grande (Diniz, 2000). O produto interno entre dois vetores iguais é chamado de *variância* e o produto interno entre dois vetores diferentes é chamado de *covariância*. Pode-se notar, então, que na diagonal principal estão as variâncias (que representam o quanto aquela dimensão varia em torno da sua média) e fora da diagonal principal estão as covariâncias (que representam o quanto as dimensões variam em torno da média uma em relação a outra). No nosso exemplo a matriz de covariância é:

$$C = \begin{pmatrix} 170{,}4556 & 139{,}3556 & 0{,}1691 \\ 139{,}3556 & 240{,}9333 & 0{,}4671 \\ 0{,}1691 & 0{,}4671 & 0{,}0025 \end{pmatrix}$$

Percebe-se que a variância da idade e do peso é muito maior que a variância da altura, e assim também a altura varia muito pouco em relação às outras dimensões.

2.1.3 Autovetores e Autovalores

Como a matriz de covariância é simétrica, por resultados de álgebra linear (Boldrini, 1980) sabemos que existe uma base ortonormal formada apenas por autovetores desta matriz e pode ser ortogonalizada da seguinte forma: $C = PDP^T$, sendo D a matriz diagonal que contém apenas os autovalores de *C* e *P* a matriz dos autovalores associados. A matriz de transformação *P* que será utilizada na equação (1) para expressar os dados em uma nova base é, na verdade, uma matriz cujas colunas são os autovetores da matriz de covariância (Richardson, 2009).

Precisamos, então, calcular os autovalores e autovetores da matriz *C*. Podemos encontrar os autovalores calculando as raízes do polinômio característico:

$$pol(\lambda) = det(C - \lambda I) = det\begin{pmatrix} 170{,}4556 - \lambda & 139{,}3556 & 0{,}1691 \\ 139{,}3556 & 240{,}9333 - \lambda & 0{,}4671 \\ 0{,}1691 & 0{,}4671 & 0{,}0025 - \lambda \end{pmatrix} = 0$$

$\lambda_1 = 0.0015$, $\lambda_2 = 61.9529$, $\lambda_3 = 349.4370$

Os autovetores, associados a cada autovalor λ_i encontrado, são obtidos resolvendo-se a equação:

$$(C - \lambda_i I)v = \begin{pmatrix} 170{,}4556 - \lambda_i & 139{,}3556 & 0{,}1691 \\ 139{,}3556 & 240{,}9333 - \lambda_i & 0{,}4671 \\ 0{,}1691 & 0{,}4671 & 0{,}0025 - \lambda_i \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = 0$$

Estes autovetores podem ainda ser normalizados, como dado abaixo:

$$v_1 = \begin{pmatrix} 0.0011 \\ -0.0026 \\ 1.0000 \end{pmatrix}, v_2 = \begin{pmatrix} -0.7890 \\ 0.6143 \\ 0.0025 \end{pmatrix}, v_3 = \begin{pmatrix} -0.6143 \\ -0.7890 \\ -0.0014 \end{pmatrix}$$

Neste processo de calcular os autovetores da matriz de covariância, somos capazes de identificas as características que melhor representam os dados através da ordenação dos autovetores de acordo com seus autovalores, do maior para o menor. Essa ordenação fica mais fácil pois a matriz é simétrica e com a diagonal positiva, o que garante autovalores reais e maiores ou iguais a zero. O que teremos é uma base na qual o primeiro autovetor indica a direção de maior variância (ou maior energia, ou maior relevância). Assim, dizemos que o autovetor associado ao maior autovalor é a principal componente do nosso conjunto de dados. O autovetor com o segundo maior autovalor é a segunda componente principal, e assim por diante. Como $\lambda_3 > \lambda_2 > \lambda_1$, v_3 é o vetor que representa a primeira componente principal, v_2é o vetor que representa a segunda componente principal e v_1a terceira componente principal.

2.1.4 Dados representados na base de autovetores

A parte final do método de análise de componentes principais corresponde a expressar os dados originais em termos da base de autovetores representada pela matriz ortogonal *P.* Tanto os dados com médias nulas armazenados na matriz *X*, quanto os dados originais da matriz *A* podem ser transformados e representados na nova base de autovetores: $Y = PA^T$ e $Z = PX^T$. Observamos que os autovetores estão dados nas linhas de *P, n* x *n.* Assim os valores de *Y* (e de forma análoga os de *Z*) são expressos em termos de projeções nas direções dos autovetores já ordenados:

$$Y = PA^T = \begin{pmatrix} - & v_3 & - \\ - & v_2 & - \\ - & v_1 & - \end{pmatrix} \begin{pmatrix} | & \dots & | \\ a_1 & \dots & a_{10} \\ | & \dots & | \end{pmatrix} = \begin{pmatrix} < v_3, a_1 > & \dots & < v_3, a_{10} > \\ < v_2, a_1 > & \dots & < v_2, a_{10} > \\ < v_1, a_1 > & \dots & < v_1, a_{10} > \end{pmatrix} \quad (4)$$

Se todas as direções forem mantidas, não haverá perda de informação, e o processo final apenas transformará os dados originais para um novo sistema de coordenadas (formado pelos autovetores) com a mesma dimensão. No exemplo inicial teremos as amostras de idade, peso e altura de cada pessoa escrito na base de autovetores.

Para retornarmos às matrizes originais *A* e *X* basta multiplicarmos *Y* e *Z* pela inversa de *P*, que é a sua transposta, pois a base de vetores é ortonormal (Boldrini, 1986): $P^T Y = A^T$ e $P^T Z = X^T$.

2.1.5 Dados em um conjunto reduzido de direções

Como as direções dadas pelos autovetores podem ter relevâncias muito diferentes no sistema de representação dos dados, podemos escolher apenas as direções principais e descartar as demais. Neste caso, haverá uma compressão de dados, ou seja, haverá uma diminuição da dimensão dos dados finais a serem representados. Isso faz com que percamos informações, mas como as direções descartadas estão associadas aos autovalores de menor módulo, estatisticamente falando, não é perdido muito.

Assim, aos invés de considerarmos *P* a matriz completa de mudança de base da canônica para a base de autovetores, vamos considerar uma matriz $\tilde{P}$ apenas com as direções principais, chamada de *matriz característica* que conterá os autovetores que queremos manter. No nosso exemplo, $\tilde{P}$ terá as duas primeiras direções principais, v_3 e v_2.

$$\tilde{Y} = \tilde{P}\,A^T = \begin{pmatrix} - & v_3 & - \\ - & v_2 & - \end{pmatrix}\begin{pmatrix} | & \dots & | \\ a_1 & \dots & a_{10} \\ | & \dots & | \end{pmatrix} = \begin{pmatrix} < v_3, a_1 > & \dots & < v_3, a_{10} > \\ < v_2, a_1 > & \dots & < v_2, a_{10} > \end{pmatrix} \quad (5)$$

Agora $\tilde{P}$, *2x10* e *Y, 3x10,* não mais possuem dados com a mesma dimensão, pois encontram-se em subespaços de dimensões diferentes. No entanto, a partir de $\tilde{Y}$ ainda é possível obtermos aproximações para as matrizes de dados originais *A* e *X*,

$$\tilde{A} = \tilde{P}^T\tilde{Y} \text{ e } \tilde{X} = \tilde{P}^T\tilde{Z} \quad (6)$$

No exemplo inicial, os dados com dimensão reduzida são apresentados abaixo, dando enfoque para grandezas associadas às componentes principais do sistema que são *idade* e *peso*:

$$\tilde{Y}^T = \begin{pmatrix} 15.2157 & -2.5953 \\ 5.8315 & -1.6255 \\ -3.9077 & -9.2509 \\ 6.9817 & 17.7569 \\ 19.4258 & -2.0711 \\ -14.6048 & -3.4568 \\ 31.0023 & -3.4804 \\ -22.9349 & 0.4942 \\ -13.3703 & 9.5230 \\ -23.6393 & -5.2942 \end{pmatrix} \quad (7)$$

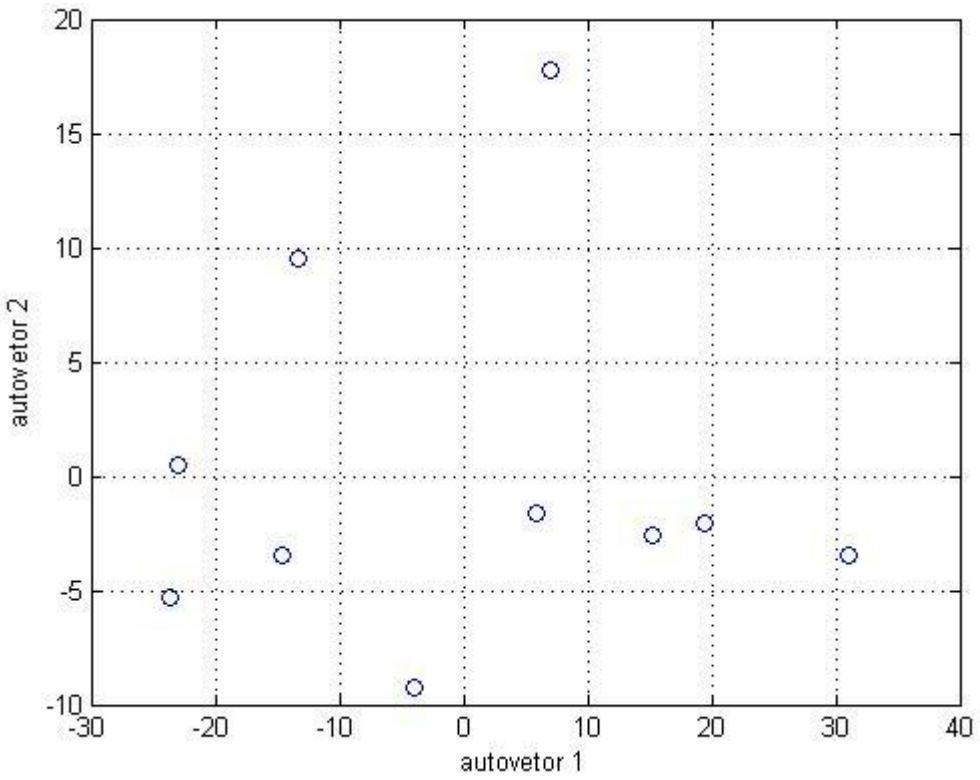

Figura 2: Dados representados na base de autovetores de dimensão 2

A Figura 2 dada a seguir mostra como os dados ficaram representados na base de autovetores que escolhemos. Observe que agora temos um subespaço de dimensão 2 dentro do R^3, já que, como identificado pela ACP, uma das direções pode ser descartada por não ser muito representativa para os dados em questão.

O que acabamos de fazer foi transformar nossos dados de forma que fiquem expressos em termos de duas características apenas, ao invés das 3 inicialmente informadas. A escolha destas 2 características foi feita de forma a melhor representar estes dados.

3. APLICAÇÕES

3.1. Compressão de imagens

Como visto na equação (6), a partir dos dados com dimensão reduzida, ainda é possível obtermos matrizes com valores aproximados $\tilde{A}$ e $\tilde{X}$, novamente com a mesma dimensão de A e X. Por essas matrizes $\tilde{A}$ e $\tilde{X}$ possuírem valores próximos dos originais e serem obtidas a partir de um conjunto com menos coordenadas, elas são muito utilizadas em aplicações que envolvam compressão de dados. Na verdade, elas são as formas comprimidas de A e X.

No processamento digital de imagens, um dado contínuo (analógico) é convertido em uma matriz de elementos simples (pixels) que assumem valores discretos (escalas de cinza). A partir dessa matriz de pixels é implementado a ACP, extraindo as componentes principais que representarão a maior parte da variância dos dados.

Um exemplo de utilização dessa técnica é na compressão de imagens médicas. A compressão reduz significativamente o espaço de memória ocupado por cada imagem. Quanto maior for a redução na dimensão de P, maior será a compressão (menos componentes principais utilizados). Na verdade, há a necessidade de buscarmos um ponto de equilíbrio para que a imagem no final deste processo não esteja degradada, ou seja, não tenha perdido informações essenciais em relação à imagem inicial. Essa ferramenta permite grande economia de espaço, que pode ser crítico em aplicações clínicas já que o volume de imagens analisadas e armazenadas pode ser bastante grande.

3.2. Separação cega de fontes (BSS – Blind Source Separation)

A principal motivação para o uso deste método é o problema conhecido como “cocktail party” (Richardson, 2009). Supondo que há N pessoas em uma festa e estão todas falando ao mesmo tempo, resultando em uma mistura de todas as vozes. O objetivo é extrair os monólogos individuais de cada pessoa. A sala contém tantos microfones quanto pessoas falando, e estes estão espalhados pela sala. Cada microfone grava uma versão diferente da mistura de vozes. Processando a ACP nesses N sinais combinados é possível retirar o ruído e separar cada sinal original separadamente.

Um uso comum desta técnica é em sinais biomédicos como o eletroencefalograma (EEG) e o eletrocardiograma (ECG). Dada a expressiva quantidade de trabalhos de separação de sinais biomédicos, pode-se dizer que esta área corresponde à principal aplicação de BSS (Duarte, 2006). Os sinais captados em métodos não-invasivos (que não inserem nenhum sensor dentro do corpo) geralmente vêm acompanhados de muito ruído devido a outras atividades fisiológicas. O fato de ser muito difícil determinar quais sinais fisiológicos interferentes são captados fornece uma boa justificativa para a aplicação de BSS. Em uma relação com o problema da “cocktail party”, cada sinal fisiológico seria equivalente a uma pessoa na festa, e o sinal captado (EEG ou ECG) seria a mistura de “vozes” captado pelo microfone.

4. CONSIDERAÇÕES FINAIS

Neste trabalho apresentamos o método de análise de componentes principais, considerada um dos resultados mais importantes para a álgebra linear aplicada (Shlens, 2003), através de sua formulação espectral, enfocando conceitos de álgebra linear e apresentando todas as etapas da formulação do método. Além disso, neste trabalho mencionamos de forma abrangente algumas das aplicações mais importantes da técnica de ACP em problemas associados a dados biomédicos.

REFERÊNCIAS

BOLDRINI; Costa; Figueiredo; Weltzler. **Álgebra Linear**. São Paulo: Harper & Row do Brasil, 1986.

DINIZ, A., **Apostila I Estatística Básica**, 2000.

DUARTE, L., **Um Estudo sobre Separação Cega de Fontes e Contribuições ao Caso de Misturas Não-lineares**, Acessado em 10 out 2014. Online. Disponível em: https://www.google.com.br/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0CCcQFjAB&url=http%3A%2F%2Fwww.bibliotecadigital.unicamp.br%2Fdocument%2F%3Fdown%3Dvtls000387543&ei=PChIVPLTK4-ONtfWgcgL&usg=AFQjCNENUdQVFZ7SktKpjNCcbrHxjT3LdQ&sig2=5Fibi3D-CnZfO5zFMniGfg&bvm=bv.77880786,d.eXY&cad=rja

RICHARDSON, M., **Principal Component Analysis**, 2009.

SHLENS, J., **A tutorial on Principal Component Analysis – Derivation, Discussion and Singular Value Decomposition**, 2003.

SMITH, L., **A tutorial on Principal Component Analysis**, 2002.

WIKIPÉDIA. **Análise de Componentes Principais**, Acessado em 06 out. 2014. Online. Disponível em: http://pt.wikipedia.org/wiki/An%C3%A1lise_de_Componentes_Principais

WIKIPÉDIA. **Covariância,** Acessado em 24 out. 2014. Online. Disponível em: http://pt.wikipedia.org/wiki/Covari%C3%A2ncia