Sobre as diferenças de aplicação entre SVD e PCA: Um estudo pragmático*

Nicollas Silva, Alan Neves, Leonardo Rocha, Fernando Mourão

DCOMP/UFSJ - São João del-Rei, MG , Brasil

{nicollasilva, aneves, lcrocha,fhmourao}@ufsj.edu.br

Abstract. *Despite the popularity of SVD and PCA, several applied studies struggle to understand and differentiate such methods. Frequently, both methods are applied without an adequate assessment about each one is the most appropriate for each scenario. In order to ease the choice between such methods in computational tasks, we conducted a pragmatic discussion that correlates the success of each application to characteristics of each evaluated domain. In this sense, we proposed a methodology that, when applied to three real collections related to distinct tasks, showed that there are differences in the application of SVD and PCA and that a not elaborate choice may be harmful to the task performed.*

Resumo. *Apesar da popularidade do SVD e PCA, há uma dificuldade comum a vários estudos aplicados em compreender e diferenciar tais métodos. Frequentemente, ambos são aplicados sem uma avaliação adequada sobre qual é o mais apropriado para cada cenário. A fim de facilitar a escolha entre tais métodos em tarefas computacionais, realizamos uma discussão pragmática que correlaciona o sucesso da aplicação destes métodos com características do domínio de análise. Para tanto, propomos uma metodologia que, aplicada em três coleções reais relacionadas a tarefas distintas, permitiu-nos verificar que há diferenças na aplicação do SVD e do PCA e que uma escolha não elaborada pode ser nociva à tarefa realizada.*

1. Introdução

Métodos numéricos oriundos da Álgebra Linear (AL) e da Estatística (ES) têm se tornado cruciais para o sucesso de variadas tarefas computacionais, tais como Classificação Automática de Documentos, Recomendação de Produtos, Recuperação de Informação, dentre outras [Eldén 2007, Wold et al. 1987, Elden 2006, Wall et al. 2003, Koren et al. 2009]. Tal relevância pode ser explicada pela robustez com que estes métodos abordam problemas computacionais comuns a diversos domínios, bem como a elegante formalização matemática imposta a tais problemas.

Dentre os vários métodos existentes em AL, destacamos o *Singular Value Decomposition* (SVD), enquanto enfatizamos o *Principal Component Analysis* (PCA), dada a popularidade e generalidade de aplicação de ambos. SVD tem por objetivo realizar a redução de posto e a aproximação de baixo-posto de uma matriz N-dimensional. Em virtude de suas propriedades algébricas, SVD é comumente aplicado visando um de três objetivos distintos: (1) eficiência de manipulação de

*Esse trabalho foi parcialmente financiado por CNPq, CAPES e Fapemig.

dados matriciais; (2) redução da dimensionalidade dos dados; (3) remoção de ruídos dos dados originais [Wall et al. 2003]. Por sua vez, o PCA é uma técnica estatística multivariada que possui como objetivo explorar a estrutura de variabilidade dos dados. De modo geral, o PCA preocupa-se em explicar a estrutura de variância-covariância de um conjunto de variáveis por meio de poucas combinações lineares entre essas variáveis. Os objetivos mais concretos dessa análise são (1) redução de dados e (2) interpretação [Johnson and Wichern 2002].

Este trabalho tem por objetivo apresentar uma discussão intuitiva e pragmática sobre as diferenças entre SVD e PCA, visando relacionar o sucesso de aplicação desses métodos com características do domínio de análise e a tarefa a ser realizada. Apesar de serem amplamente estudados por matemáticos e estatísticos, e utilizados em diversas áreas da computação, SVD e PCA são comumente confundidos quando aplicados em cenários práticos. Poucos estudos sabem quando preterir o uso de um método frente ao outro em aplicações computacionais. Salvo em estudos mais teóricos, em geral, os conceitos do SVD e PCA são apresentados de forma confusa e a escolha por um deles não é sequer discutida. Perguntas tais como: *Quais as diferenças práticas entre SVD e PCA? A escolha não elaborada entre estes métodos pode degradar a qualidade de uma tarefa? Quando utilizar o SVD ou o PCA?* apesar de pertinentes são negligenciadas na literatura. Responder tais perguntas permitiria-nos alcançar mais eficientemente resultados com melhor qualidade.

A fim de alcançar nossos objetivos, realizamos manipulações sobre provas matemáticas de otimalidade de ambas técnicas. Assim, identificamos que: (1) O SVD foca na descoberta de "identidades"[1] predominantes dos dados; (2) o PCA foca na descoberta de "distorções"[2] marcantes nos dados. A partir de tais observações, propomos uma metodologia de análise dos dados útil para suportar a escolha entre tais métodos de acordo com características de médias e variações apresentadas pelas variáveis de cada domínio. Aplicando a metodologia proposta em três coleções reais, verificamos que a distinção entre SVD e PCA não é devidamente realizada, mesmo em cenários clássicos. Por exemplo, verificamos que embora a grande maioria dos estudos adotem o PCA para a tarefa de reconhecimento facial, o SVD, apontado por nossa metodologia como mais apropriado, apresentou um ganho de 60% sobre um tradicional *benchmark* para esta tarefa. Ressaltamos que as contribuições deste trabalho são particularmente relevantes para estudos aplicados, uma vez que aspectos teóricos são pouco discutidos e quando o são, ocorrem de maneira pouco pragmática e relevante para as análises almejadas. Cabe ainda salientar que não encontramos na literatura trabalhos que abordem os aspectos levantados sobre o uso e distinção de SVD e PCA de maneira similar à proposta neste trabalho.

Enfatizamos que todas as implementações, execuções de experimentos e análises de resultados foram realizadas pelo aluno Nicollas Silva, sob a orientação dos professores Leonardo Rocha e Fernando Mourão. A análise teórica sobre os métodos foi conduzida em conjunto, aluno e professores. Além disso, esse trabalho contou com a colaboração do aluno Alan Neves na concepção da metodologia proposta.

[1]Definimos como identidade de uma variável sua média amostral.

[2]Definimos como distorção de uma variável sua variância amostral a partir de sua identidade.

2. Conceitos Básicos & Trabalhos Relacionados

SVD e PCA têm sido extensivamente aplicados a variadas áreas da computação. Mineração de Dados, Aprendizado de Máquina, Recuperação de Informação, Processamento Digital de Sinais, dentre outras, encontram nestes métodos uma maneira eficaz de abordar algumas de suas principais tarefas [Wall et al. 2003, Elden 2006]. Grande parte destes trabalhos aplicados, entretanto, não discutem a decisão de se utilizar uma técnica ou outra. Além disso, discussões teóricas, quando existentes, são confusas ou pouco pragmáticas, dificultando a escolha acertada entre ambos métodos, em cada tipo de domínio.

SVD consiste em um processo de fatoração de matrizes capaz de representar uma matriz de dados A por meio de três outras matrizes U, S e V^T que representam, respectivamente: uma base ortonormal[3] para as colunas de A (i.e., autovetores a esquerda); o conjunto de escalares que determinam a relevância de cada autovetor (i.e., autovalores); e uma base ortonormal para as linhas de A (i.e., autovetores a direita). Tais bases representam um restrito conjunto de dimensões independentes capazes de gerar toda a informação contida na matriz A. Além disso, o SVD é capaz ainda de "ordenar" a informação contida em A, tornando a "parte dominante" visível, uma vez que os autovetores estão ordenados decrescentemente pela relevância definida pelos autovalores [Eldén 2007]. Dessa forma, através do SVD é possível excluir informações redundantes da matriz A. Isso é possível graças a uma propriedade da AL, que garante que a combinação linear dos vetores base é capaz, e suficiente, de gerar todos os vetores do espaço em questão. O SVD permite ainda descartarmos informações pouco discriminativas da matriz A, obtendo a melhor aproximação possível para A usando um número pequeno de dimensões distintas.

O PCA, por sua vez, é uma técnica estatística multivariada que possui como objetivo explorar a estrutura de variabilidade dos dados. Os principais conceitos estatísticos para entender o processo são: *média*, *variância* e *covariância*. A *média* pode ser definida como uma medida que busca sintetizar a informação de tendência central da distribuição de valores da variável. A *variância* é uma medida de dispersão estatística que indica quão longe os valores de uma variável se encontram da média. Já a *covariância*, serve para medir o grau de relacionamento linear entre duas variáveis. Segundo [Johnson and Wichern 2002], o objetivo do PCA é explicar a estrutura de variância e covariância entre variáveis através da construção de poucas combinações lineares das variáveis originais, e o que se deseja obter é a "**redução** do número de variáveis a serem avaliadas e a **interpretação** das combinações lineares construídas". Dessa forma, a informação contida nas variáveis originais é substituída pela informação contida nos k ($k \leq min(m, n)$) componentes principais não correlacionados. De acordo com [Smith 2002], as etapas para aplicação do PCA consistem em:

1. *Escolher variáveis (i.e., dimensões) a serem avaliadas*;
2. *Subtrair média de cada dimensão, produzindo um conjunto de dados com média zero*;
3. *Calcular matriz de covariância*;
4. *Calcular os autovetores e autovalores da matriz de covariância*;
5. *Escolher as k componentes principais(i.e., os k autovetores com maior autovalor)*;
6. *Interpretar as informações contidas nas componentes principais.*

[3]Conjunto de vetores, com norma igual a 1, linearmente independentes, capazes de gerarem todos os outros vetores de A.

Diversos trabalhos têm substituído a etapa 4 pela aplicação do SVD (PCA/SVD), uma vez que os K primeiros autovetores da matriz de covariância são as k dimensões de maior variabilidade. Neste sentido, podemos interpretar o PCA como o processo de fatoração do SVD, aplicado à matriz de covariância. Devido a isso, alguns autores consideram ambos métodos como equivalentes [Abdi 2007], embora outros ainda os consideram como diferentes, uma vez que a análise do PCA não é necessariamente realizada por meio do SVD [Johnson and Wichern 2002]. Assim, salvo as devidas diferenças teóricas, a questão prática se reduz à identificarmos as diferenças entre SVD sobre a matriz de covariância e SVD sobre a matriz original.

3. SVD × PCA: Análise Teórica

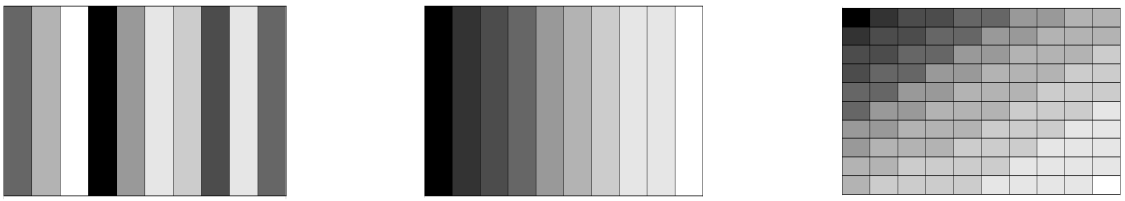
Nesta seção, provaremos teoricamente a distinção SVD e PCA. Para tanto, baseamo-nos na prova de que os componentes principais identificados pelo PCA minimizam o erro de aproximação para uma dada matriz com média corrigida (i.e., matriz com colunas cuja média é zero), apresentada por [Johnson and Wichern 2002].

Seja X a matriz de dados; Z a matriz resultante do produto vetorial XX'; X_0 a matriz de média corrigida; e Σ a matriz de covariância sobre a matriz X. Em [Johnson and Wichern 2002], os autores demonstram que o objetivo do PCA pode ser reduzido ao problema de determinar os autovetores da matriz Σ, uma vez que cada um desses autovetores apontam na direção de menor erro residual de X_0. De maneira análoga, reduz-se o objetivo do SVD a determinar os autovetores da matriz Z, dado que cada autovetor aponta na direção de menor erro residual de X.

Dessa forma, entender as diferenças entre PCA e SVD consiste em entendermos as diferenças entre os autovetores das matrizes Z e Σ. Com este intuito, considere as matrizes X e X_0 novamente. Se colorirmos as colunas de X a fim de representar a média de cada dimensão, em que quanto mais escura a cor, maior é o valor absoluto da média, temos uma matriz com aspecto similar ao apresentado na Figura 1 (a), em que observamos um aspecto não ordenado das variáveis. O mesmo pode ser dito se representarmos cada coluna da matriz X_0 a partir de seu valor. Entretanto, neste caso, os valores se referem à variância. Suponha agora que ordenemos as colunas da matriz X decrescentemente pela média, tal como ilustrado na Figura 1 (b). Ao realizar o mesmo processo para a matriz X_0 temos uma ordenação decrescente pela variância. Por questões de nomenclatura, chamaremos estas matrizes ordenadas de X^H e X_0^H, respectivamente. Suponha agora que geremos as matriz:

$$Z^H = X^H X^{H'} \text{ e } \Sigma^H = \frac{1}{n-1} X_0^H X_0^{H'}$$

As matrizes Z^H e Σ^H teriam um aspecto similar ao apresentado pela Figura 1 (c). Porém, novamente enquanto as cores mais escuras representam variáveis com maior média para a matriz Z^H, referem-se às variáveis com maior variância e covariância para a matriz Σ^H. É importante ressaltar que as matrizes Z e Z^H são equivalentes, visto que a única diferença entre elas está na ordenação relativa das colunas e linhas. Porém, para o processo de decomposição e descoberta de autovetores tais matrizes geram os mesmos autovetores e autovalores, apenas com valores em ordem distintas [McWorter and Meyers 1998]. O mesmo pode ser dito sobre as matrizes Σ e Σ^H.



(a) Variáveis não ordenadas

(b) Ordenação decrescente

(c) Produto escalar sobre variáveis ordenadas

Figura 1. Representação de Matrizes coluna.

Considerando o objetivo do SVD, a direção de menor erro residual é o primeiro autovetor u_1 de Z e maximiza o produto $u_1' Z u_1$. Observando nossa representação da matriz Z^H, percebemos que u_1, obrigatoriamente, priorizará sua componente associada a primeira coluna da matriz Z^H. Para ver isso, vamos usar uma prova por contradição. Suponha que $u_{1,i}$ (i.e., a i-ésima componente de u_1) seja sua componente de maior valor absoluto com $i \neq 1$. Neste caso, existe um outro vetor w_1 que possui as mesmas componentes que u_1 salvo que $w_{1,1} = u_{1,i}$ e $w_{1,i} = u_{1,1}$. Neste caso, o produto $w_1' Z^H w_1$ geraria um valor maior que $u_1' Z^H u_1$ e, consequentemente, u_1 não seria o vetor que maximiza $u_1' Z^H u_1$ (ou equivalentemente $u_1' Z u_1$). Como u_1 é um vetor unitário e $u_1' Z u_1$ deve ser máximo, seus maiores coeficientes devem estar associados às variáveis com maiores valores de Z^H. Assim, no caso da matriz Z as variáveis (dimensões) com maiores médias contribuem mais para definir os autovetores. Considerando a matriz Σ^H, o raciocínio é o mesmo. Mas, neste caso, as dimensões com maior variabilidade contribuem mais para definir os autovetores.

Dessa forma, concluímos que os autovetores do SVD tendem a priorizar dimensões com médias maiores, valorizando as identidades predominantes da coleção. Por outro lado, os autovetores do PCA apontam para as dimensões de maior variabilidade, valorizando as distorções predominantes. Uma pergunta pertinente que surge é: porque SVD e PCA apresentam resultados similares em muitos casos? Sabe-se que variáveis com maiores médias tendem a apresentar uma maior variabilidade em cenários reais. Neste caso, a ordenação relativa das variáveis na matriz X^H se torna igual a observada na matriz X_0^H e, consequentemente a matriz Z^H se torna igual a matriz Σ^H. Com isso, os autovetores de Σ e Z tendem a apontar para a mesma direção e o resultado se torna similar.

4. SVD × PCA: Metodologia de Análise

De forma a validar empiricamente as distinções entre SVD e PCA, discutidas na Seção 3, propomos uma metodologia de análise sobre estes métodos. Tal metodologia estabelece métricas e procedimentos que nos permitem correlacionar características inerentes às coleções de entrada com indicadores de qualidade na aplicação de cada método. Dessa forma, seremos capazes de diferenciar o desempenho do SVD e PCA em domínios distintos, identificando qual a técnica mais apropriada em cada caso.

Definimos como características das coleções, qualquer informação que sumarize numericamente um comportamento relacionado aos valores presentes em uma amostra de dados. Por exemplo, o número de dimensões, a distribuição de médias por dimensão, ou mesmo uma matriz de covariância são características que podemos derivar de amostras. Por questões de eficiência, estamos interessados apenas

em características cujo custo computacional de cálculo seja menor que o custo de se executar o SVD ou PCA. Por sua vez, indicadores de qualidade são informações que permitem-nos contrastar a matriz resultante após se aplicar o SVD ou PCA com a matriz de dados original. Observe que a relevância de cada indicador de qualidade, neste caso, depende do objetivo final de análise do estudo que utiliza SVD ou PCA como técnicas de tratamento dos dados.

Tabela 1. Características e Indicadores de Qualidade utilizados na metodologia.

Caraterísticas da Coleção	1.**Distribuição de *Ranking* (AUC-DR):** Para cada dimensão i, primeiramente, calculamos a média dos valores de i e a covariância média da dimensão i com as demais. Em seguida, normalizamos as médias e covariâncias médias e ordenamos decrescentemente as dimensões por estes valores, gerando duas distribuições de rank distintas. Por fim, geramos a AUC (*Area Under the Curve*) para cada distribuição. Quanto maior a AUC, mais dimensões relevantes a coleção possui. 2.**Correlação de *Ranking* (CR):** Geramos duas distribuições de rank para as dimensões de uma coleção. Uma distribuição baseada na média de cada dimensão e outra baseada na covariância média entre as dimensões. Por fim, calcula-se o coeficiente de correlação *kendall-tau* entre as duas distribuições. Quanto maior este coeficiente, mais similares são as informações de média e covariância nos dados. 3.**Vizinhança por Similaridade (AUC-VS):** Para cada par de dimensões da coleção, calculamos a similaridade entre elas utilizando a distância Cosseno de seus respectivos vetores. Em seguida, definimos como vizinhas quaisquer dimensões i e j cuja similaridade S_{ij} seja maior que um limiar τ. Posteriormente, variando o valor de τ em 0.1 no intervalo de 0 a 1, calculamos o número médio de vizinhos que cada dimensão possui. Em seguida, plotamos o tamanho médio de vizinhança por τ. Por fim, calculamos a AUC sob a curva gerada. Quanto maior a AUC, mais similaridades, ou redundâncias, há entre as dimensões da coleção.
Indicadores de Qualidade	1.**Erro de Aproximação Total (EAT):** Calcula-se a soma do erro absoluto entre cada posição da matriz original e a correspondente posição na aproximação de posto k, obtida pelo SVD ou PCA. 2.**Erro de Aproximação Médio (EAMe):** Primeiro, obtém-se o erro absoluto médio por dimensão entre a matriz original e sua aproximação de posto K. Por fim, calcula-se a média dos erros obtidos por dimensão. 3.**Erro de Aproximação Máximo (PAMx):** Primeiro, obtém-se o erro absoluto médio por dimensão entre a matriz original e sua aproximação de posto K. Por fim, identifica-se o maior erro absoluto médio. 4.**Distribuição de Erros (AUC-DE):** Calcula-se o erro absoluto entre cada posição da matriz original e a correspondente posição na aproximação de posto k. Em seguida, ordena-se decrescentemente tais erros e gera-se uma distribuição de rank. Por fim, calcula-se a AUC para a distribuição gerada. 5.**Homogeneidade de Erros (AUC-HE):** Calcula-se o erro absoluto entre cada posição da matriz original e a correspondente posição na aproximação de posto k. Em seguida, para cada dimensão define-se a diferença entre o maior e o menor erro. Ordena-se decrescentemente tais diferenças e gera-se uma distribuição de rank. Por fim, calcula-se a AUC para a distribuição gerada.

A metodologia proposta consiste na execução de cinco etapas: (1) Definir um conjunto de características de interesse sobre as coleções de entrada; (2) Definir um conjunto de indicadores de qualidade aderente a distintos objetivos de análise; (3) Derivar as características em domínios reais; (4) Rodar o SVD e o PCA sobre dados reais e aplicarmos os indicadores selecionados sobre os resultados; (5) Analisar correlações entre os valores de características e indicadores encontrados sobre as amostras reais. O intuito é correlacionar comportamentos de características com o comportamento dos indicadores em distintos cenários, de forma a identificar, ou aprender, padrões relevantes e recorrentes que nos permitam prever o comportamento dos indicadores baseado apenas nas características. A Tabela 1 apresenta as características e indicadores de qualidade analisados neste trabalho. Tais características e indicadores foram levantados considerando as distinções teóricas discutidas na Seção 3, bem como os objetivos comumente relacionados a ambos métodos na literatura. Dada a complexidade de nossas análises, restringimo-nos, neste trabalho, apenas a verificar a existência de tais padrões em importantes cenários de aplicações do SVD e PCA. A definição de um projeto experimental apropriado, que nos permitirá validar cada padrão encontrado representa uma importante direção de pesquisa futura. Cabe ainda salientar que nossa metodologia baseia-se em um conjunto não fechado de características e indicadores de qualidade que podemos derivar de cenários reais de aplicação. De fato, a medida de que novos cenários forem avaliados, novas características e indicadores devem ser incluídos na metodologia.

5. Estudos de Caso

Nesta seção, avaliamos a metodologia proposta em três cenários reais que veem na aplicação do SVD e PCA uma transformação sobre os dados necessária para se atingir seus objetivos finais. Neste intuito, descrevemos cada um desses cenários, bem como as bases de dados utilizadas. Posteriormente, discutimos os resultados de se aplicar nossa metodologia em tais coleções de dados.

5.1. Cenários de Análise & Bases de Dados

O primeiro cenário é o de Reconhecimento Facial (RF), que consiste em, dado uma nova foto bidimensional da face de um indivíduo, identificar se este está presente em um conjunto de fotos já processadas. O desafio de tratamento neste caso deve-se a grande quantidade de imagens que podem estar indexadas nos bancos de dados. Portanto, técnicas de redução de dimensionalidade proporcionadas pelo SVD e PCA são de grande contribuição. Estudos sobre RF, usualmente, adotam o PCA neste processo sem, porém, darem maiores explicações sobre essa escolha. Selecionamos como base para nossas análises a Yale Face Database[4]. Esta base contém 165 imagens de 15 indivíduos, no formato GIF. Há 11 fotos em escala de cinza, por indivíduo, cada uma contendo diferentes expressões faciais e configurações de ambiente.

O segundo cenário é o de Recuperação de Informação (RI). Em RI, a premissa comum é que existem poucos tópicos semânticos latentes na comunicação humana [Deerwester et al. 1990]. Baseado nessa premissa, almeja-se extrair tais tópicos semânticos de um corpo de texto analisando-se as associações entre os termos que ocorrem em contextos semelhantes. No denominado método LSI (*Latent Semantic Analysis*), o SVD é comumente aplicado sobre as coleções e cada autovetor resultante é visto como um tópico latente. Embora mais recentemente outras técnicas vêm sendo aplicadas neste processo, tais como NMF e LDA [Wallach 2006], para tarefas tradicionais de RI, como classificação de documentos, SVD e PCA ainda são amplamente utilizados e selecionados sem critérios bem definidos. A coleção de dados referente a este cenário é uma coleção de notícias publicadas pela Reuters e contém 8.144 documentos, organizados em 8 classes, e 24.985 termos distintos.

Por fim, analisamos o cenário de Bioinformática (BF). Neste caso, é comum o uso do PCA ou SVD como etapa de pré-processamento dos dados para agrupamento de expressões gênicas, por exemplo, [Lan et al. 2003]. O objetivo é, muitas vezes, remover os ruídos ou variações inerentes a dados genéticos, que não impactam o resultado esperado da análise. Novamente, PCA e SVD são utilizados para essa finalidade sem uma escolha apropriada. Para analisar este cenário, adotamos uma base de taxas de expressão para 6.118 genes da *Saccharomyces cerevisiae*, coletados em 7 momentos temporais distintos durante o processo de esporulação [Raychaudhuri et al. 2000].

5.2. Análise de Qualidade

Analisando as características de cada cenário, tal como apresentado na Tabela 2, observamos que as informações de média e covariância são altamente correlacionadas em todos os casos (i.e., possuem CR ≥ 0.80). Porém, observamos algumas diferenças

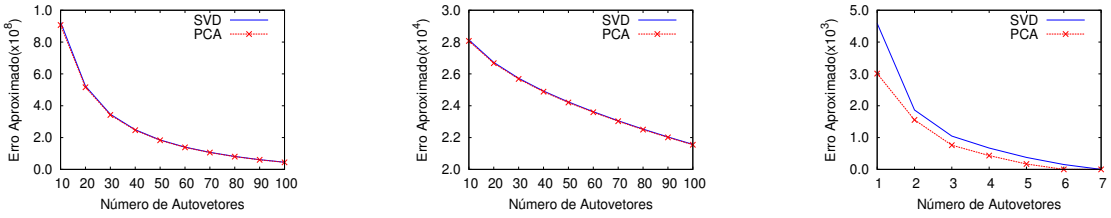
[4]cvc.yale.edu/projects/yalefaces/yalefaces.html

entre estes tipos de informação. Nos cenários RF e BF, as médias apresentam uma quantidade maior de informação relevante que a covariância, exibindo um AUC-DR maior. Observamos também que o cenário com maior nível de informações redundantes é o RF, com AUC-VS igual a 0.74. Além disso, observamos que o cenário BF possui poucas dimensões distintas relevantes, apresentando um AUC-DR menor que 0.1.

Tabela 2. Análise de características das coleções analisadas.

Cenário de análise	AUC-DR		CR	AUC-VS
	Média	Covariância		
RF	0.4142	0.2479	0.9297	0.7489
RI	0.2871	0.3314	0.8710	0.1177
BF	0.0415	0.0361	0.8207	0.5918

Considerando os indicadores de qualidade, por restrição de espaço, focaremos nossas discussões no EAT e na AUC-DE. A Figura 2 apresenta os EAT gerados para distintos números de autovetores. Observamos que o cenário com menor correlação entre SVD e PCA possui a maior diferença, quanto a erro de aproximação, entre o uso do SVD e PCA. Os outros dois cenários apresentam resultados bem similares, apesar das distinções discutidas. Porém, uma análise mais detalhada sobre estes erros nos aponta algumas diferenças importantes entre o uso do SVD e PCA. A Figura 3 apresenta a métrica AUC-DE para diferentes números de autovetores. Para o cenário RF, observamos que o SVD concentra erros maiores em poucas dimensões distintas, gerando uma AUC-DE menor. A medida que o número de autovetores aumenta, o erro do SVD, embora como um todo diminua se espalhando entre mais dimensões, torna-se pior que o PCA. Dessa forma, o uso do SVD com poucos autovetores é a melhor solução neste caso. A análise de homogeneidade corrobora essa conclusão, apresentando comportamento similar ao gráfico da Figura 3. Por sua vez, no cenário RI as diferenças entre SVD e PCA permanecem pequenas, tanto considerando a AUC-DE quanto os demais indicadores de qualidade. Isso sugere que o uso do SVD e PCA sobre a amostra estudada teriam comportamentos similares.



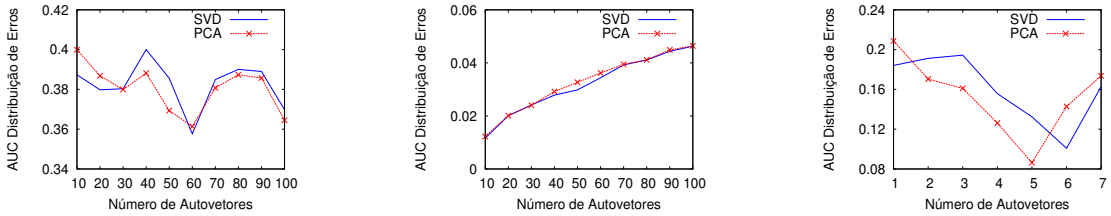
(a) Cenário RF (b) Cenário RI (c) Cenário BF

Figura 2. Análise de Erros Aproximados Totais (EAT).

5.3. Análise de Impacto

Os resultados discutidos na Seção 5.2 suscitam uma questão primordial: qual o impacto dos erros associados ao SVD e PCA sobre a análise final em cada cenário? Por restrição de espaço, limitamos a discussão sobre essa questão ao cenário de RF.

A fim de ressaltar o impacto do SVD e PCA sobre o reconhecimento facial, simulamos um sistema tal como segue. Primeiro, dado um conjunto de P fotos,



(a) Cenário RF (b) Cenário RI (c) Cenário BF

Figura 3. Análise de Distribuição de Erros (AUC-DE).

selecionamos aleatoriamente uma foto como teste e as $P-1$ restantes como treino. A partir deste conjunto de treino, construímos uma matriz, $trainMatrix$, que contém a representação vetorial de cada foto. Ou seja, cada coluna de $trainMatrix$ consiste em uma foto f_i distinta e cada linha representa uma célula de f_i. Dessa forma, considerando que cada foto bidimensional f_i tenha dimensões $M \times N$, temos $M \times N$ linhas distintas em $trainMatrix$ e $P-1$ colunas distintas. Posteriormente, aplicamos o SVD e o PCA sobre a matriz $trainMatrix$, de forma a obter a melhor aproximação pela combinação dos K primeiros autovetores ($K << \min(M \times N, P)$). De posse desta representação K-dimensional, a definição de um sistema de identificação consiste, primeiro, em representar uma foto de consulta f_t, com dimensões $M \times N$, no novo espaço K-dimensional. Em seguida, identificar qual das fotos de treino mais se aproxima de f_t. Especificamente, utilizamos a distância euclidiana entre as coordenadas no espaço K-dimensional para definir similaridade entre fotos. A foto f_i com menor distância é retornada como a mais provável de representar a mesma pessoa presente na foto f_t.

A Figura 4 ilustra a diferença de desempenho ao se aplicar SVD e PCA com diferentes valores de K. Observamos que o SVD é capaz de representar nitidamente fotos usando apenas 40 autovetores, que representa 60% menos autovetores que o PCA necessita para obter uma nitidez similar. Este comportamento foi observado em 10 repetições aleatórias deste experimento. Além de uma melhor representação, o uso do SVD gera uma grande economia de espaço para modelar e armazenar fotos neste tipo de aplicação. Dessa forma, apesar de aparentemente pequenas, as diferenças de erros entre SVD e PCA trazem um impacto significativo para o reconhecimento facial.

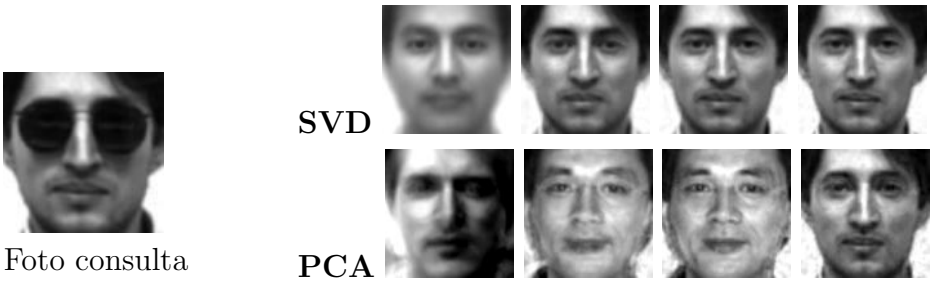


Figura 4. Comparação entre SVD e PCA sob a tarefa de reconhecimento facial.

6. Conclusões e Trabalhos Futuros

Neste trabalho, propomos uma metodologia de análise que estabelece métricas e procedimentos para correlacionar características inerentes às coleções de entrada com

indicadores de qualidade na aplicação do SVD e PCA. Tal metodologia, visa identificar, em distintos cenários de aplicação, padrões relevantes e recorrentes que nos possibilitam prever o comportamento dos indicadores baseado apenas nas características. De fato, avaliações empíricas em três cenários reais, bem estabelecidos na literatura, permitiu-nos verificar a existência de alguns desses padrões, bem como o impacto de se realizar uma escolha não elaborada entre SVD e PCA. A definição de um projeto experimental apropriado, que nos permitirá validar cada padrão encontrado, representa uma direção de pesquisa futura. Por fim, ressaltamos a consolidação de uma ferramenta que permita auxiliar pesquisadores na tomada de decisão entre SVD e PCA, de acordo com a aplicação, como principal trabalho futuro e contribuição dessa pesquisa.

Referências

Abdi, H. (2007). Singular value decomposition (svd) and generalized singular value decomposition.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407.

Elden, L. (2006). Numerical linear algebra in data mining. *Acta Numerica*, 15, s. 327-384.

Eldén, L. (2007). *Matrix methods in data mining and pattern recognition*. Society for Industrial and Applied Mathematics.

Johnson, R. and Wichern, D. (2002). *Applied multivariate statistical analysis*, volume 4. Prentice Hall Upper Saddle River, NJ.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Lan, H., Stoehr, J. P., Nadler, S. T., Schueler, K. L., Yandell, B. S., and Attie, A. D. (2003). Dimension reduction for mapping mrna abundance as quantitative traits. *Genetics*, 164(4):1607–1614.

McWorter, W. A. and Meyers, L. F. (1998). Computing eigenvalues and eigenvectors without determinants. *Mathematics magazine*, pages 24–33.

Raychaudhuri, S., Stuart, J. M., and Altman, R. B. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. In *Pacific Symposium on Biocomputing*, page 455. NIH Public Access.

Smith, L. I. (2002). A tutorial on principal components analysis.

Wall, M. E., Rechtsteiner, A., and Rocha, L. M. (2003). *Singular Value Decomposition and Principal Component Analysis*, chapter 5, pages 91–109. Kluwel.

Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.