



## Análise de Componentes Principais (PCA)

### Exercício

No exercício proposto, foram obtidos dados contendo 400 fotos dos rostos de 40 pessoas diferentes, sendo 10 de cada, em posições ligeiramente diferentes. Cada foto possui 4096 pixels (64x64), sendo que cada pixel pode assumir um valor de 0 a 255, correspondendo a uma escala de preto e branco. Assim, cada pixel possui 1 byte de tamanho, e cada foto aproximadamente 4KB, sendo a base inteira composta por 800KB de dados. A biblioteca que gerou estas imagens é a *RnavGraphImageData* do R.

As fotos foram consideradas dados de entrada e foram divididas em 4 *folds*, seguindo então a proporção percentual de 75/25 em relação aos dados para treinamento e teste, respectivamente. Cada pixel tornou-se então uma variável diferente do problema.

Devido ao alto número de variáveis envolvidas, a técnica PCA foi implementada para definir as variáveis que possuíam maior variância nos dados. Assim, mesmo que com uma certa perda de valores, foi possível analisar apenas uma quantidade considerada pequena de pixels para definir a quem os rostos do grupo de teste pertenciam. Para isso, cada pessoa foi considerada uma classe e, com isso, o problema tornou-se um classificador para definir em qual das 40 classes (pessoas) um certo dado de teste (foto) pertencia.

Para a técnica PCA, primeiro foram calculados os autovalores e os autovetores utilizando funções próprias do R. Na figura 1 são mostrados os maiores autovalores, em ordem decrescente, das variáveis. As 3 primeiras variáveis correspondem a aproximadamente 46% da variância total do problema, enquanto as 8 primeiras correspondem a aproximadamente 70%. Porém, para garantir uma boa acurácia da classificação, foram utilizados dados que correspondiam a mais de 98% da variância total do problema.

Foram feitas 10 simulações, dividindo-se os dados em 75/25 para cada iteração. O número médio de pixels analisados nas divisões foi de 109, uma redução de pouco menos de 36 vezes na quantidade de valores. Assim, foram analisados e testados apenas 43KB em média de dados, contra os 800KB totais.

Quanto à acurácia, a média dos testes foi de 75,96% e o desvio padrão dos resultados foi de 5,05%. A tabela das acurácias obtidas em cada iteração é mostrada na tabela 1. Já a matriz de confusão é mostrada na tabela 2.

Iteração	Acurácia
1	82,83%
2	72%
3	80,20%
4	72%
5	74%
6	67%
7	76,24%
8	75,25%
9	82,83%
10	77,23%
Média	75,96%

Tabela 1

Mesmo utilizando as variáveis correspondentes a 98% da variância total do problema, a acurácia média manteve-se baixa em termos relativos. Isso se explica pois ao dividir o conjunto de fotos em dados para o teste e para o treinamento, as fotos com alta variação (como quando uma pessoa tirou fotos com óculos e sem óculos, ou quando uma pessoa afastou-se demais da câmera ou mesmo inclinou demais a cabeça

