3 Métodos

3.1 Análise de Componentes Principais

A Análise de Componentes Principais (PCA - Principal Component Analysis) (48, 49) é um método estatístico que visa reduzir a dimensão de problemas com muitas medidas. O método PCA elimina redundâncias e transforma um sistema descrito por um conjunto de variáveis possivelmente correlacionadas em um novo sistema descorrelacionado. A orientação dos eixos no espaço original é alterada de forma que os eixos de maior dispersão dos dados se tornem os novos eixos de referência espacial.

Os dados podem ser organizados como uma matriz W de dimensões $N \times M$, onde as N linhas são as observações e as M colunas representam as medidas. Como PCA é sensível às variações de escala dos dados, é importante primeiro realizar a normalização dos dados. Cada elemento da linha i na coluna j é subtraído pela média e dividido pelo desvio padrão desta coluna:

$$X_{ij} = \frac{W_{ij} - \bar{w}_j}{s_j} \ , \tag{3.1}$$

em que

$$\bar{w}_j = \frac{\sum_{i=1}^{N} W_{ij}}{N} \ , e \tag{3.2}$$

$$s_j = \sqrt{\frac{\sum_{i=1}^{N} (W_{ij} - \bar{w}_j)^2}{N-1}} \tag{3.3}$$

são respectivamente a média e o desvio padrão da medida j. O próximo passo é calcular a matriz de covariância V:

$$V_{ij} = \frac{\sum_{k=1}^{N} (X_{ki} - \bar{x}_i)(X_{kj} - \bar{x}_j)}{N-1} \ , \tag{3.4}$$

em que $\bar{x}_i$ é a média da medida i. Deve-se ressaltar que como os dados foram previamente normalizados, a média é sempre igual a zero ($\bar{x}_i = 0$). Em seguida, são calculados os au-

tovalores λ e autovetores $\vec{v}_\lambda$ de V. Cada autovalor está associado a um autovetor. Assim, ao organizar os autovalores em ordem decrescente, os autovetores também são respectivamente ordenados. Esta ordenação é realizada porque quanto maior o autovalor, maior a quantidade de variação dos dados é explicada pela componente associada. Desta forma, para realizar redução de dimensionalidade minimizando as perdas de informação, basta selecionar os primeiros P valores, de acordo com quantas dimensões se pretende gerar. Portanto, os autovetores $\vec{v}_\lambda$ são ordenados de acordo com os autovalores λ, resultando na nova matriz de autovetores $\vec{p}_\lambda$. O próximo passo é realizar a transformação linear:

$$W' = (\vec{p}_\lambda X)^T \ . \tag{3.5}$$

A quantidade de variação dos dados explicada pelos P autovetores escolhidos pode ser quantificada pela seguinte expressão:

$$r = \frac{\sum_{j=1}^{P} \lambda_j}{\sum_{j=1}^{M} \lambda_j} \ . \tag{3.6}$$

3.2 Análise Canônica

A Análise Canônica (50–52), também conhecida como análise de discriminantes lineares (*Linear Discriminant Analysis* - LDA), é um método que busca encontrar a projeção que melhor separa classes de dados pré-definidas. Isto é obtido a partir da maximização da dispersão inter-classe, ou seja dispersão entre classes, enquanto minimiza a dispersão intra-classe dentro de cada classe. Considerando que cada elemento da matriz normalizada X, definida na Equação 3.1, pode ser classificado em uma classe C_i contendo n_i elementos, em que $i = 1, 2, ..., N_c$ e N_c é o número máximo de classes. Assim, as matrizes de dispersão inter-classe (Equação 3.7) e intra-classe (Equação 3.8) podem ser definidas como:

variação total

Média de Ci

$$S_{\text{inter}} = \sum_{i=1}^{N_c} n_i (\langle \vec{x} \rangle_i - \langle \vec{x} \rangle)(\langle \vec{x} \rangle_i - \langle \vec{x} \rangle)^T \ , \tag{3.7}$$

numero elementos de Ci

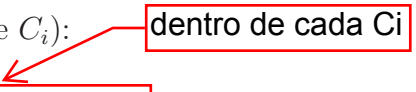
Média de todos

$$S_{\text{intra}} = \sum_{i=1}^{N_c} S_i \ , \tag{3.8}$$

variação total das classes

em que $\langle \vec{x} \rangle_i$ é o vetor de características médio dos elementos na classe C_i, $\langle \vec{x} \rangle$ é o vetor de características médio de todos os elementos, e S_i é a dispersão das medidas dentro de

cada classe (matriz de dispersão para cada classe C_i):



$$S_i = \sum_{k \in C_i} (\vec{x_k} - \langle \vec{x} \rangle_i)(\vec{x_k} - \langle \vec{x} \rangle_i)^T . \tag{3.9}$$

Desta forma, pode-se finalmente calcular os autovalores e autovetores da matriz $S_{\text{intra}}^{-1} S_{\text{inter}}$, em que S_{intra}^{-1} é a inversa de S_{intra}. Em seguida, os autovalores devem ser organizados em ordem decrescente. Assim, podem-se selecionar os autovetores correspondentes aos maiores autovalores para a nova projeção, com dimensionalidade reduzida.