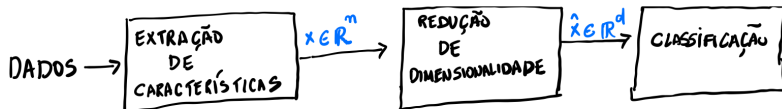


Análise de Componentes Principais

Advanced Institute for Artificial Intelligence – AI2

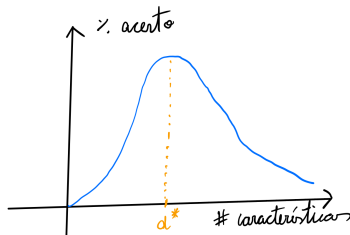
<https://advancedinstitute.ai>

Técnicas de redução de dimensionalidade/transformação do espaço de características visam obter versões mais **compactas/representativas** de nossos dados. Dado um conjunto de dados $\mathcal{X} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_z, y_z)\}$ em que $\mathbf{x}_i \in \mathbb{R}^n$ e $y_i \in \mathbb{N}$, a ideia consiste em obter um novo conjunto $\hat{\mathcal{X}} = \{(\hat{\mathbf{x}}_1, y_1), (\hat{\mathbf{x}}_2, y_2), \dots, (\hat{\mathbf{x}}_z, y_z)\}$, em que $\hat{\mathbf{x}}_i \in \mathbb{R}^d$ e $d < n$.

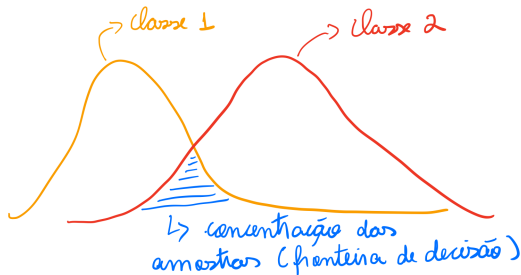


Usualmente, em problemas de classificação, temos a impressão de, quanto mais características temos, melhor será a taxa de acerto de nossa técnica. No entanto, em problemas reais em que temos um número **limitado de amostras**, observa-se um fenômeno conhecido por **maldição da dimensionalidade**. Existe, então, uma série de efeitos negativos ocasionados pelo aumento indiscriminado de características:

- Fenômeno de Hughes: para um número finito de amostras, existe uma dimensionalidade d^* que, após este valor, o desempenho da taxa de classificação diminui.



- Número de amostras como função das características: em classificadores não paramétricos, o número de amostras deve ser uma função exponencial do número de características.
- Gaussianas multivariadas: em distribuições Gaussianas multivariadas com alta dimensão, a densidade das amostras tende a se concentrar na cauda da distribuição, ou seja, longe da média amostral, dificultando a classificação.



Autovalores e Autovetores: Uma Breve Introdução

Seja V um espaço vetorial com n dimensões que contempla um produto interno e dois vetores $\mathbf{u}, \mathbf{v} \in V$. Supondo que V seja um espaço Euclidiano, temos que seus vetores possuem uma **direção** e uma **magnitude**. O **produto interno** entre seus vetores pode ser calculado da seguinte forma:

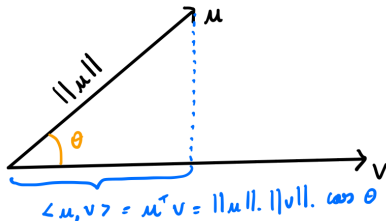
$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T \mathbf{v} = \sum_{i=1}^n u_i v_i. \quad (1)$$

Uma outra possibilidade para calcular o produto interno entre dois vetores é dada como segue:

$$\langle \mathbf{u}, \mathbf{v} \rangle = \|\mathbf{u}\| \cdot \|\mathbf{v}\| \cdot \cos(\theta), \quad (2)$$

em que θ corresponde ao ângulo entre os dois vetores e $\|\cdot\|$ representa a magnitude do vetor.

Esta formulação modela o produto interno entre os vetores u e v como sendo a **projeção** de u em v .



Temos, também, que:

$$\langle v, v \rangle = v^T v = \sum_{i=1}^n v_i^2 = \|v\|^2. \quad (3)$$

Existem diferentes maneiras de calcular a norma (magnitude) de um vetor. A norma Euclidiana $\|\mathbf{v}\|_2$ de um vetor, também chamada de norma L_2 , é calculada da seguinte forma:

$$\|\mathbf{v}\|_2 = \sqrt{\sum_{i=1}^n v_i^2}. \quad (4)$$

Usualmente, denotamos a norma Euclidiana por $\|\cdot\|$.

A ideia é que neste espaço vetorial eu consiga definir **operadores lineares**, ou seja, funções (matrizes) \mathbf{P} que mapeiam vetores de entrada \mathbf{v} em vetores de saída \mathbf{u} , ou seja:

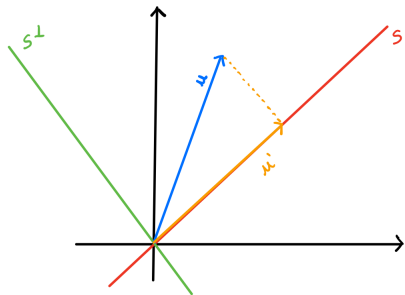
$$\mathbf{u} = \mathbf{P}\mathbf{v}. \quad (5)$$

No estudo de autovalores e autovetores, estamos interessados em operadores lineares que possuem a seguinte característica: dado um operador linear P para quais vetores v a saída da Equação 3, ou seja, u , aponta para a mesma direção da entrada, sendo apenas esticados ou encolhidos? Matematicamente falando, temos que:

$$u = Pv = \lambda v. \tag{6}$$

Todos os vetores v que satisfazem a equação acima são chamados de **autovetores** de P , e todos os escalares λ que satisfazem a formulação acima são chamados de **autovalores** de P . Na prática, λ é um fator de escala de redução ou incremento do vetor v . Assim sendo, dizemos que $v \in \mathbb{R}^n$ é um autovetor de $P \in \mathbb{R}^{n \times n}$ com autovalor λ se $v \neq 0$ e $Pv = \lambda v$.

Para fins de explicação, tomemos o seguinte exemplo: seja $\mathbf{P} = \mathbf{I} \in \mathbb{R}^{n \times n}$ o operador identidade. Então, $\mathbf{I}\mathbf{v} = 1\mathbf{v}$, $\forall \mathbf{v} \in \mathbb{R}^n$. Desta forma, \mathbf{v} é um autovetor de \mathbf{I} com autovalor $\lambda = 1$. Vejamos um outro exemplo, o operador de projeção em \mathbb{R}^2 .



Podemos definir o operador \mathbf{P}_S , o qual projeta o vetor \mathbf{u} no subespaço (reta) S :

$$\mathbf{u}' = \mathbf{P}_S \mathbf{u}. \quad (7)$$

Supondo um vetor $\mathbf{w} \in S$, temos que $\mathbf{w} = \mathbf{P}_S \mathbf{w}$, dado que \mathbf{w} já faz parte do subespaço S . Neste caso, todo vetor que pertence à reta S é um autovetor de \mathbf{P}_S com autovalor $\lambda = 1$.

Seja, agora, S^\perp o subespaço ortogonal a S , e $\mathbf{x} \in S^\perp$, ou seja, quando eu aplicar o operador \mathbf{P}_S em algum elemento de S^\perp , resulta no valor 0 (origem). Então, temos que $\mathbf{P}_S \mathbf{x} = 0\mathbf{x}$, ou seja, todo vetor que pertence a S^\perp é um autovetor de \mathbf{P}_S com autovalor $\lambda = 0$.

Podemos extrapolar esse arcabouço de autovetores e autovalores para matrizes. Por exemplo, como podemos calcular os autovalores e autovetores de uma matriz \mathbf{A} ? Da Equação 6, temos que:

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}.$$

Rearranjando os termos, temos:

$$\mathbf{A}\mathbf{v} - \lambda\mathbf{I}\mathbf{v} = 0 \implies (\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = 0. \quad (8)$$

Existe um teorema da álgebra linear que nos diz o seguinte: $\mathbf{B}\mathbf{v} = 0$ admite solução não nula se, e somente se, $\det(\mathbf{B}) = 0$. Assim, $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ em nosso caso.

Ex: seja a matriz:

$$\mathbf{A} = \begin{bmatrix} 3/2 & -1/2 \\ -1/2 & 3/2 \end{bmatrix}.$$

Temos que:

$$\begin{aligned} \mathbf{A} - \lambda \mathbf{I} &= \begin{bmatrix} 3/2 & -1/2 \\ -1/2 & 3/2 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 3/2 - \lambda & -1/2 \\ -1/2 & 3/2 - \lambda \end{bmatrix}. \end{aligned}$$

Sabemos que $\det(\mathbf{A} - \lambda \mathbf{I}) = 0$, ou seja:

$$\left(\frac{3}{2} - \lambda\right)^2 - \left(-\frac{1}{2} * -\frac{1}{2}\right) = \left(\frac{3}{2} - \lambda\right)^2 - \frac{1}{4} = 0,$$

o que implica em:

$$\frac{9}{4} - 2\frac{3}{2}\lambda + \lambda^2 - \frac{1}{4} = 0 \implies \lambda^2 - 3\lambda + 2 = 0.$$

Assim sendo, temos uma equação do segundo grau cujas soluções são $\lambda_1 = 1$ e $\lambda_2 = 2$.

Para obtermos o autovetor associado ao autovalor $\lambda_1 = 1$, temos que:

$$\begin{aligned} \mathbf{A} - \lambda \mathbf{I} &= \begin{bmatrix} 3/2 & -1/2 \\ -1/2 & 3/2 \end{bmatrix} - 1 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 3/2 - 1 & -1/2 \\ -1/2 & 3/2 - 1 \end{bmatrix} \\ &= \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix}. \end{aligned} \tag{9}$$

Continuando, substituindo o resultado da Equação 7 na Equação 6, temos:

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{v} = 0$$

$$\begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix} \mathbf{v} = 0$$

$$\begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \implies \begin{cases} \frac{1}{2}v_1 - \frac{1}{2}v_2 = 0 \\ -\frac{1}{2}v_1 + \frac{1}{2}v_2 = 0 \end{cases} \implies v_1 = v_2 \implies \mathbf{v}^{(1)}$$

Desta forma, $\mathbf{v}^{(1)}$ é o autovetor associado ao autovalor λ_1 .

Note que temos infinitos autovetores associados ao autovalor λ_1 , basta apenas que as componentes tenham o mesmo valor. Ex: $\mathbf{v}^{(1)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ ou $\mathbf{v}^{(1)} = \begin{bmatrix} 7 \\ 7 \end{bmatrix}$. Qual a diferença entre eles?

Todos apontam para a mesma direção, mas possuem magnitudes diferentes.

Usualmente, utilizamos o **autovetor canônico**, ou seja, aquele que tem **norma unitária**. Basta tomarmos um autovetor qualquer e dividirmos cada componente pela sua norma, ou seja:

$$\mathbf{v}^{(1)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \implies \mathbf{v}^{(1)} = \begin{bmatrix} \frac{1}{\|\mathbf{v}^{(1)}\|} \\ 1 \end{bmatrix} \implies \mathbf{v}^{(1)} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 1 \end{bmatrix}.$$

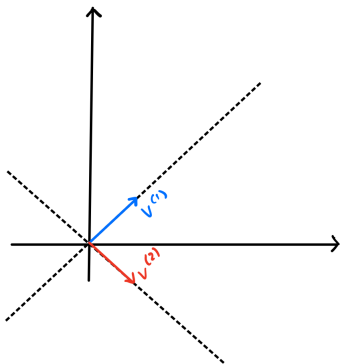
Agora, repetimos o processo para o autovalor $\lambda_2 = 2$:

$$\begin{bmatrix} -1/2 & -1/2 \\ -1/2 & -1/2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \implies \begin{cases} -\frac{1}{2}v_1 - \frac{1}{2}v_2 = 0 \\ -\frac{1}{2}v_1 - \frac{1}{2}v_2 = 0 \end{cases} \implies v_1 = -v_2 \implies \mathbf{v}^{(2)}.$$

Novamente, utilizamos o autovetor canônico:

$$\mathbf{v}^{(2)} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \implies \mathbf{v}^{(2)} = \begin{bmatrix} \frac{1}{\|\mathbf{v}^{(1)}\|} \\ 1 \\ -\frac{1}{\|\mathbf{v}^{(1)}\|} \end{bmatrix} \implies \mathbf{v}^{(2)} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 1 \\ -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

Desta forma, temos que $v^{(1)}$ e $v^{(2)}$ são os nossos autovetores do operador linear P_S . Geometricamente falando, temos o seguinte:



Temos que os autovetores definem uma nova base no sistema de coordenadas. Existe um teorema que nos diz o seguinte: “Se P é um operador linear que possui n autovalores distintos, então os autovetores de P definem uma nova base em \mathbb{R}^n . ”

Podemos organizar os autovetores em uma matriz Q da seguinte forma:

$$Q = \begin{bmatrix} \mathbf{v}^{(1)} & \mathbf{v}^{(2)} & \dots & \mathbf{v}^{(n)} \end{bmatrix}$$

Ademais, temos que uma matriz é dita ser **positiva semidefinida** se todos os seus autovalores foram maiores ou iguais a 0. Temos, ainda, uma outra definição: "Seja P um operador linear. Caso a matriz Q dos seus autovetores defina uma nova base em \mathbb{R}^n , então $PQ = Q\Lambda$, em que Λ representa a matriz diagonal dos autovalores".

Um outro teorema fundamental (decomposição expectral) nos diz que: "Seja P uma matriz quadrada $n \times n$, Q a sua matriz de autovetores (nas colunas) e Λ a sua matriz diagonal de autovalores. Então, temos que a matriz P pode ser decomposta da seguinte forma:"

$$P = Q\Lambda Q^{-1}. \tag{10}$$

Caso P seja ortogonal (ex: matrizes de rotação do espaço), então $Q^{-1} = Q^T$. Desta forma, temos que $P = Q\Lambda Q^T$.

A técnica PCA nos permite tratar dados de alta dimensionalidade identificando a dependência entre as variáveis para representá-los de uma forma mais compacta e minimizando a perda de informação relevante. É uma das técnicas mais utilizadas no contexto de **redução de características**. Possui outros nomes, tais como: (i) Transformação de Karhunen-Loeve, (ii) Transformação de Hotelling e (iii) Decomposição em Valores Singulares.

Características principais:

- Transformação linear: assume a hipótese que os dados encontram-se em um subespaço Euclidiano de \mathbb{R}^n .
- Método não supervisionado.
- Decorrelaciona os dados de entrada eliminando redundâncias (a matriz de covariância dos dados transformados é diagonal).

$$\mathbf{x} \in \mathbb{R}^n \xRightarrow{\text{PCA}} \hat{\mathbf{x}} \in \mathbb{R}^d$$

Basicamente, queremos achar um operador que projete os dados de entrada em um espaço de saída com menor dimensão.

Seja $Z = [T^T, S^T]$ uma base ortonormal em \mathbb{R}^n , em que:

- $T^T = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d]$ corresponde ao vetor de componentes que iremos **reter** no processo de redução e
- $S^T = [\mathbf{w}_{k+1}, \mathbf{w}_{k+2}, \dots, \mathbf{w}_n]$ corresponde ao vetor de componentes que iremos **descartar**.

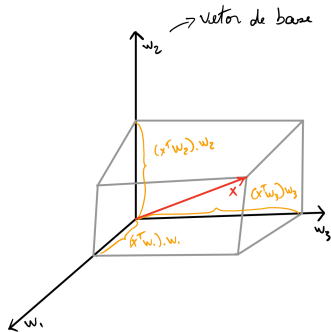
Desta forma, T representa o novo sistema de eixos coordenados encontrados pelo PCA, enquanto que S representa o subespaço eliminado durante o processo de redução.

Problema: dado um espaço de entrada, queremos encontrar as d direções \mathbf{w}_i que, ao projetarmos os dados, maximizam a variância (espalhamento) retida na nova representação.

Temos que nosso vetor $\mathbf{x} \in \mathbb{R}^n$ pode ser expandido em nossa base ortonormal da seguinte forma:

$$\mathbf{x} = \sum_{j=1}^n (\mathbf{x}^T \mathbf{w}_j) \mathbf{w}_j = \sum_{j=1}^n c_j \mathbf{w}_j, \quad (11)$$

em que $\mathbf{c} \in \mathbb{R}^n$ corresponde ao vetor de coeficientes de expansão.



Note que o termo $\mathbf{x}^T \mathbf{w}_j$ corresponde à projeção de \mathbf{x} no vetor de base \mathbf{w}_j .

Temos que o vetor reduzido para o novo espaço pode ser calculado da seguinte forma:

$$\hat{\mathbf{x}} = T\mathbf{x} \implies \hat{\mathbf{x}}^T = \mathbf{x}^T T^T = \sum_{j=1}^n c_j \mathbf{w}_j^T T^T. \quad (12)$$

Note que a propriedade de ortonormalidade (bases ortonormais) nos diz que:

$$\mathbf{w}_i^T \mathbf{w}_j = \begin{cases} 1 & \text{se } i = j \\ 0 & \text{caso contrário.} \end{cases} \quad (13)$$

Assim, podemos reescrever a Equação 10 conforme segue:

$$\hat{\mathbf{x}} = \sum_{j=1}^n c_j \mathbf{w}_j^T T^T = \sum_{j=1}^n c_j \mathbf{w}_j^T [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d] = [c_1, c_2, \dots, c_d]. \quad (14)$$

Assim, queremos encontrar uma transformação T (conjunto de componentes) que **maximize** a variância retida nos dados, ou seja, queremos maximizar a seguinte função:

$$L(T) = E \left[\|\hat{\mathbf{x}}\|^2 \right] = [\hat{\mathbf{x}}^T \hat{\mathbf{x}}] = \sum_{j=1}^d E[c_j^2], \quad (15)$$

lembrando que $\|\hat{\mathbf{x}}\|^2$ corresponde à norma ao quadrado de $\hat{\mathbf{x}}$, ou seja, queremos obter novas componentes cujas normas (distâncias a partir da origem) sejam maximizadas, ou seja, estejam mais “espalhadas”. Note que $[\hat{\mathbf{x}}^T \hat{\mathbf{x}}] = E[\hat{x}_1^2] + E[\hat{x}_2^2] + \dots + E[\hat{x}_d^2]$, ou seja, a soma das variâncias em cada eixo coordenado da nova representação.

Da Equação 9, temos que $c_j = \mathbf{x}^T \mathbf{w}_j$. Assim, substituindo-se c_j na Equação 15, temos que:

$$\begin{aligned} L(T) &= \sum_{i=1}^d E[c_i^2] = \sum_{i=1}^d E[(\mathbf{x}^T \mathbf{w}_i)^2] \\ &= \sum_{i=1}^d E[\mathbf{w}_i^T \mathbf{x} \mathbf{x}^T \mathbf{w}_i] = \sum_{i=1}^d \mathbf{w}_i^T E[\mathbf{x} \mathbf{x}^T] \mathbf{w}_i \\ &= \sum_{i=1}^d \mathbf{w}_i^T \Sigma_{\mathbf{x}} \mathbf{w}_i, \end{aligned} \tag{16}$$

em que $\Sigma_{\mathbf{x}}$ corresponde à matriz de covariância dos dados observados, ou seja, do vetor de entrada \mathbf{x} . Ademais, temos que a Equação 16 está sujeita à restrição $\|\mathbf{w}_i\|^2 = 1$.

Desta forma, o problema de otimização consiste em:

$$\mathbf{w}^* = \arg \max_{\mathbf{w}} \sum_{i=1}^d \mathbf{w}_i^T \Sigma_x \mathbf{w}_i, \quad (17)$$

sujeito à $\|\mathbf{w}_i\|^2 = 1, \forall i = 1, 2, \dots, d$. Desta forma, nosso problema consiste em encontrar as d direções ortogonais \mathbf{w}_i que maximizam o somatório acima. Note que o somatório contém a matriz de covariância dos dados, ou seja, queremos maximizar a sua variância nessas direções ortogonais.

Temos, então, um problema de otimização com restrições de desigualdade, ou seja, precisamos fazer uso da técnica de Multiplicadores de Lagrange.

Escrevendo a Equação 16 de acordo com os Multiplicadores de Lagrange, temos que:

$$L(T, \alpha) = \sum_{i=1}^d \mathbf{w}_i^T \Sigma_{\mathbf{x}} \mathbf{w}_i - \sum_{i=1}^d \alpha_i \underbrace{(\mathbf{w}_i^T \mathbf{w}_i - 1)}_{\|\mathbf{w}\|^2=1}. \quad (18)$$

Como queremos maximizar a equação acima visando cada \mathbf{w}_i , basta, então, calcularmos a derivada de $L(T, \alpha)$ em relação à \mathbf{w}_i e igualarmos à 0, ou seja:

$$\frac{\partial L(T, \alpha)}{\partial \mathbf{w}_i} = \Sigma_{\mathbf{x}} \mathbf{w}_i - \alpha_i \mathbf{w}_i = 0 \implies \Sigma_{\mathbf{x}} \mathbf{w}_i = \alpha_i \mathbf{w}_i. \quad (19)$$

O resultado da equação acima nos leva à formulação do problema de encontrar os autovalores e autovetores. **Assim, as direções \mathbf{w}_i são os autovetores da matriz de covariância dos dados de entrada.**

Podemos reescrever a Equação 17 da seguinte forma:

$$\begin{aligned} \mathbf{w}^* &= \arg \max_{\mathbf{w}} \sum_{i=1}^d \mathbf{w}_i^T \Sigma_{\mathbf{x}} \mathbf{w}_i = \arg \max_{\mathbf{w}} \sum_{i=1}^d \mathbf{w}_i^T \alpha_i \mathbf{w}_i \\ &= \arg \max_{\mathbf{w}} \sum_{i=1}^d \alpha_i \cancel{\|\mathbf{w}_i\|^2}^1 = \arg \max_{\mathbf{w}} \sum_{i=1}^d \alpha_i. \end{aligned} \quad (20)$$

A equação acima nos diz que devemos maximizar o somatório dos d autovalores. Isto define o nosso problema, ou seja, encontrar as d direções que maximizam a variância dos dados!

Agora, veremos por que PCA decorrelaciona os dados, ou seja, a matriz de covariância do vetor transformado \hat{x} é uma matriz diagonal. Suponha a seguinte transformação **linear** abaixo:

$$\hat{x} = Ax. \quad (21)$$

Existe um teorema que diz a matriz de covariância de \hat{x} pode ser obtida da seguinte forma:

$$\Sigma_{\hat{x}} = A^T \Sigma_x A. \quad (22)$$

Ademais, temos que toda matriz simétrica e positiva semidefinida possui uma decomposição $A = Q\Lambda Q^T$, em que Q é a matriz coluna dos autovetores e Λ é a matriz diagonal dos seus autovalores (**decomposição espectral**, como vimos anteriormente). **Sabemos que as matrizes de covariância são positivas e semidefinidas!**

Desta forma, podemos reescrever a Equação 22 da seguinte forma:

$$\Sigma_{\hat{x}} = \mathbf{A}^T \Sigma_x \mathbf{A} = \mathbf{A}^T \underbrace{\mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T}_{\Sigma_x} \mathbf{A}. \quad (23)$$

Note que a matriz \mathbf{A} na Equação 21, caso consideremos essa transformação linear sendo dada pelo PCA, acaba sendo a própria matriz de autovetores, ou seja, $\mathbf{A} = \mathbf{Q}$. Assim, podemos reescrever a Equação 23 como segue:

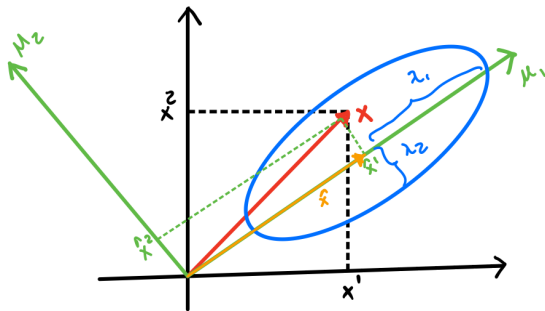
$$\Sigma_{\hat{x}} = \mathbf{A}^T \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T \mathbf{A} = \mathbf{Q}^T \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T \mathbf{Q}. \quad (24)$$

Como a nossa base é **ortonormal**, temos que $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}$. Assim, chegamos à uma formulação final para $\Sigma_{\hat{\mathbf{x}}}$:

$$\Sigma_{\hat{\mathbf{x}}} = \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & \lambda_d \end{bmatrix}, \quad (25)$$

que é uma **matriz diagonal**, ou seja os dados estão decorrelacionados! Assim, após a transformação PCA, dizemos que não há correlação entre os atributos selecionados.

Vejam, agora, uma interpretação geométrica do PCA.



Neste caso, temos que os eixos coordenados na cor preta representam o espaço original, e os eixos verdes representam o espaço transformado via PCA. Note que os novos eixos são **ortogonais** entre si. O vetor x acaba sendo projetado no eixo u_2 , pois a variância nele é maior do que em u_1 .

Um outro ponto interessante é que os autovalores λ_1 e λ_2 são, na verdade, as **variâncias dos novos eixos**, ou seja, eles codificam o quanto alongada está a elipse.

Análise de Componentes Principais pela Minimização do Erro Quadrático Médio

Um segundo critério para formularmos o problema de otimização do PCA é pelo erro quadrático médio. Desta forma, o PCA é ótimo em dois sentidos, isto é, ele **maximiza** o espalhamento dos dados e também **minimiza** o erro quadrático médio do vetor original para o vetor reduzido. Assim, podemos reformar a função a ser otimizada para:

$$L(T) = E \left[\|\mathbf{x} - \hat{\mathbf{x}}\|^2 \right] = E \left[\left\| \mathbf{x} - \sum_{j=1}^d (\mathbf{w}_j^T \mathbf{x}) \mathbf{w}_j \right\|^2 \right]. \quad (26)$$

Aplicando a definição de norma ao quadrado, temos que:

$$L(T) = E \left[\|\mathbf{x} - \hat{\mathbf{x}}\|^2 \right] = E \left[\left(\mathbf{x} - \sum_{j=1}^d (\mathbf{w}_j^T \mathbf{x}) \mathbf{w}_j \right) \left(\mathbf{x} - \sum_{j=1}^d (\mathbf{w}_j^T \mathbf{x}) \mathbf{w}_j \right)^T \right]. \quad (27)$$

Após aplicarmos operações distributivas e rearranjando os termos, temos que:

$$L(T) = E \left[\|\mathbf{x}\|^2 \right] - \sum_{j=1}^d E \left[(\mathbf{w}_j^T \mathbf{x})^2 \right]. \quad (28)$$

Note que o primeiro termo não depende de \mathbf{w}_j (que é o que queremos encontrar), sendo uma constante. Desta forma, para minimizar o erro quadrático médio, precisamos **maximizar** o segundo termo, pois ele é negativo. Este termo nada mais é do que a variância retida em cada novo eixo coordenado da base T .

Assim, podemos reescrever a Equação 28 como segue:

$$\begin{aligned} L(T) &= E \left[\|\mathbf{x}\|^2 \right] - \sum_{j=1}^d E \left[(\mathbf{w}_j^T \mathbf{x})^2 \right] \\ &= E \left[\|\mathbf{x}\|^2 \right] - \sum_{j=1}^d E \left[(\mathbf{w}_j^T \mathbf{x} \mathbf{x}^T \mathbf{w}_j) \right] \\ &= E \left[\|\mathbf{x}\|^2 \right] - \sum_{j=1}^d \mathbf{w}_j^T E \left[\mathbf{x} \mathbf{x}^T \right] \mathbf{w}_j \\ &= E \left[\|\mathbf{x}\|^2 \right] - \sum_{j=1}^d \mathbf{w}_j^T \Sigma_{\mathbf{x}} \mathbf{w}_j \end{aligned} \tag{29}$$

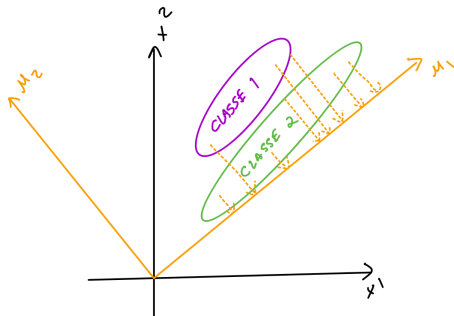
A equação anterior nos remete à mesma forma quadrática da formulação do PCA com a maximização das variâncias. **Assim sendo, para minimizarmos o erro quadrático médio, basta maximizarmos o espalhamento (variâncias) novamente!** Como fizemos anteriormente, basta derivarmos a função de custo dada pela Equação 29 com relação à \mathbf{w}_j , $\forall j = 1, 2, \dots, d$ e igualarmos à 0. Chegaremos no mesmo caso anterior, ou seja, a solução é dada na forma de uma equação de autovalores e autovetores, isto é, $\Sigma_{\mathbf{x}} \mathbf{w}_j = \lambda_j \mathbf{w}_j$. Isto significa que, voltando à Equação 29, temos a seguinte equivalência:

$$\begin{aligned} L(T) &= E \left[\|\mathbf{x}\|^2 \right] - \sum_{j=1}^d \mathbf{w}_j^T \Sigma_{\mathbf{x}} \mathbf{w}_j \\ &= E \left[\|\mathbf{x}\|^2 \right] - \sum_{j=1}^d \mathbf{w}_j^T \lambda_j \mathbf{w}_j = E \left[\|\mathbf{x}\|^2 \right] - \sum_{j=1}^d \lambda_j. \end{aligned} \quad (30)$$

O último termo é simplificado por conta das propriedade de ortonormalidade.

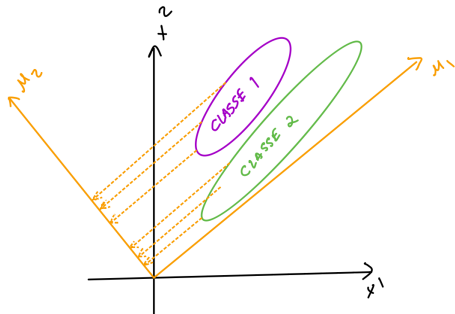
Algumas limitações do PCA:

- PCA foi projeto para **compactação** de dados, e não classificação. Caso a gente queira compactar os dados, PCA é a primeira escolha pois ele é ótimo neste sentido. No entanto, ele não é supervisionado.



Podemos observar que, nesta situação, PCA vai projetar as amostras na direção de maior variância, ou seja, u_1 . No entanto, as amostras das classes diferentes ficarão todas sobrepostas.

No entanto, caso projetássemos as amostras na direção u_2 (melhor espalhamento), elas estariam melhor separadas. Esta é, então, uma limitação da técnica PCA.



Vejam, então, o algoritmo do PCA.

- ➊ Carregar a base de dados $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$, tal que $\mathbf{x}_i \in \mathbb{R}^n$.
- ➋ Calcule o vetor média global $\boldsymbol{\mu}_x \in \mathbb{R}^n$ e a matriz de covariância $\Sigma_x \in \mathbb{R}^{n \times n}$.
- ➌ Calcule os autovetores e autovalores de Σ_x .
- ➍ Selecione os d autovetores associados aos d maiores autovalores da matriz de covariância, denotados por $[\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d]$.
- ➎ Defina a matriz de transformação $\mathbf{W}_{PCA} \in \mathbb{R}^{n \times d}$.
- ➏ Projetar dados na nova base $\hat{\mathbf{x}}_i = \mathbf{W}_{PCA}^T \mathbf{x}_i, \forall i = 1, 2, \dots, m$.