PCA

Matheus do Ó Santos Tiburcio

7 de dezembro de 2024

Resumo

Primeiro revisamos o problema de alta dimensionalidade nos dados e como reduzir a dimensão pode mostrar padrões que em dimensões maiores não eram tão claros. Em seguida analisamos mais de perto o PCA (*Principal Component Analysis*), uma das diversas técnicas de redução de dimensionalidade, para o caso de redução para o $\mathbb{R}^1$.

1 Alta dimensionalidade

É bem comum que os dados com que trabalhamos possuam dimensões altíssimas. Entretanto muitas vezes analisar dados em altas dimensões costuma ser algo bem mais difícil. Nessas dimensões os dados costumam ser esparsos (problema conhecido como *maldição da dimensionalidade*) e tratar e analisá-los é computacionalmente mais caro e até intratável dependendo da dimensão. Para um exemplo, imagine uma imagem 1080×1920, uma resolução bem comum atualmente. Se tratarmos cada posição desta imagem como uma coordenada em nossa dimensão, estaríamos em dimensão 2.073.600. Porém na representação **RGB**, que é uma das mais utilizadas, cada pixel possui 3 valores referentes aos 3 canais de cores **vermelho**, **verde** e **azul**. Dessa forma, a dimensão final resultante seria $3 \times 2.073.600 = 6.220.800$. Uma alternativa para casos assim seria reduzir a dimensão do dado tentando não o distorcer muito no processo. Em outras palavras, reduzir sua dimensão buscando perder o mínimo de informação possível.

A pergunta natural de surgir é "Como reduzir a dimensão dos dados?"e "O que seria preservar informação?". E a verdade é que nenhuma destas perguntas possui somente uma resposta! O problema é definido de maneira solta propositalmente. Essa noção de perda de informação e o modo como queremos que nossos dados reduzidos fiquem podem variar de problema para problema.

De fato, existem técnicas e técnicas para redução de dimensionalidade, lineares e não-lineares. A que focaremos aqui será o **PCA** (*Principal Component Analysis*). Apesar de sua simplicidade e característica linear, o **PCA** é amplamente utilizado e explicita relações bem interessantes entre os dados. O **PCA** busca reduzir a dimensão partindo de um pressuposto que veremos mais adiante.

2 PCA e seu pressuposto

Imagine que tenhamos vários pontos distribuídos em $\mathbb{R}^2$ como mostra a Figura 1. É notável que os pontos estão bem dispersos na direção da reta r, mas não se distanciam tanto dela. A ideia é que se preservássemos somente a informação de o quão dispersos nesta direção os dados estão, manteríamos boa parte da informação. Isto é, estão tão próximos da reta e variam tanto na direção dela, que tomar somente a direção dela parece ser uma boa representação dos dados em $\mathbb{R}^1$.

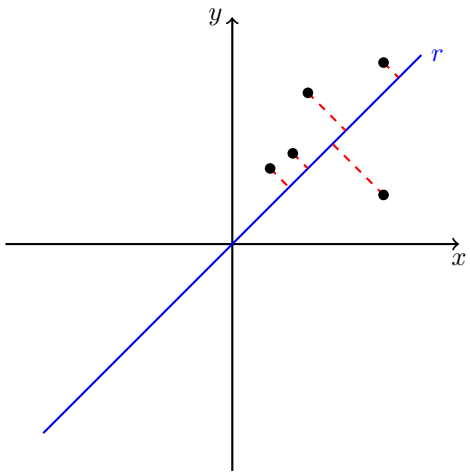

Figura 1: Gráfico de uma reta r e cinco pontos com suas distâncias para a reta representadas com linhas tracejadas.

Podemos tratar essas distâncias dos pontos para a reta, representadas por linhas tracejadas na Figura 1, como a perda de informação que teríamos se jogássemos todos os dados para a reta. A dúvida que pode surgir agora é se há alguma outra reta r que faça essas distâncias serem menores. O pressuposto do **PCA** é exatamente esse: Encontrar retas, planos e outros espaços de modo que minimizem essas distâncias. Com esses espaços encontrados, o **PCA** projetará ortogonalmente todos os dados nele.

Mas note que isso ainda está em aberto. O que significa "minimizar as distâncias?", temos mais de uma. Minimizamos uma a uma? Colocamos mais pesos na distância de alguns pontos em comparativo com outros? O **PCA** toma uma decisão bem simples e direta: Minimizaremos a soma dos erros ao quadrado. O "quadrado"é para facilitar o cálculo da otimização. Precisaremos derivar em um dado ponto e usar o quadrado das distâncias torna essa derivação mais simples.

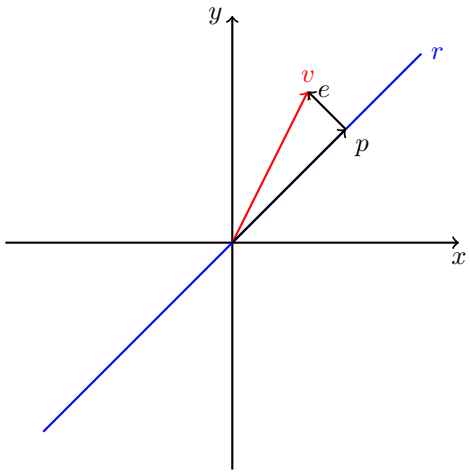

Figura 2: Gráfico de uma reta r, um vetor v, sua projeção na reta p e o vetor de erro e.

Para melhor visualização, vamos focar em redução de dimensão de $\mathbb{R}^n$ para $\mathbb{R}^1$ e imaginar os pontos como sendo vetores. Na Figura 2 se tem o vetor v e sua projeção p na reta r. Podemos imaginar que nosso erro será exatamente a norma do vetor e por seu comprimento ser exatamente a distância de v para a reta. Como queremos o erro ao quadrado, podemos tratar o erro de v como sendo $\|e\|^2$. Agora como encontrar e? Uma característica interessante é o fato de que estamos projetando ortogonalmente na reta r, então o ângulo entre o vetor e e a reta é de 90° graus. Desse modo v, p e e formam um triângulo retângulo e $\|e\|^2$ pode ser encontrado usando o teorema de pitágoras.

$$\begin{aligned} \|v\|^2 &= \|p\|^2 + \|e\|^2 \\ \|e\|^2 &= \|v\|^2 - \|p\|^2 \end{aligned} \tag{1}$$

$\|v\|^2$ já conhecemos, resta saber $\|p\|^2$. Analisando o triângulo da Figura 3 mais detalhadamente,

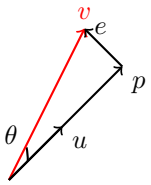

Figura 3: Triângulo retângulo formado pelos vetores v, p e e.

conseguimos deduzir $\|p\|$ usando a definição de cosseno.

$$\cos(\theta) = \frac{\text{cateto adjacente}}{\text{hipotenusa}} = \frac{\|p\|}{\|v\|} \tag{2}$$

segue que

$$\|p\| = \|v\| \cos(\theta) \tag{3}$$

Sabemos também que dado dois vetores v e u, o cosseno do ângulo entre eles assume a forma

$$cos(v, u) = \frac{v^T u}{\|v\|\|u\|} \tag{4}$$

Então podemos reescrever o cosseno da Equação (3) usando a definição (4) usando v e qualquer vetor na reta. Essa última afirmação vale pelo fato de qualquer vetor na reta r ter exatamente ângulo θ com v.

$$\begin{aligned} \|p\| &= \|v\| \frac{v^T u}{\|v\|\|u\|} \\ &= \frac{v^T u}{\|u\|} \end{aligned} \tag{5}$$

Reescrevendo $\|e\|$ com o que temos

$$\|e\|^2 = \|v\|^2 - \|p\|^2 = \|v\|^2 - \left(\frac{v^T u}{\|u\|}\right)^2 \tag{6}$$

Com essas discussões conseguimos escrever o que queremos minimizar. Queremos minimizar a soma dos erros de cada ponto ao quadrado. Assumindo que temos n pontos $x_1, x_2, ..., x_n$, queremos minimizar

$$E(u) = \|e_1\|^2 + \|e_2\|^2 + ... + \|e_n\|^2 \tag{7}$$

Isto é, queremos um vetor u cujo o erro da projeção dos pontos na reta em que ele está seja a menor possível. Agora veremos como fazer isso.

3 Minimizando $E(u)$

Como queremos o vetor u e não o valor de $E(u)$, buscaremos o **argumento mínimo** da função. Em outras palavras, buscaremos u que resulta no menor valor de $E(u)$. O problema fica

$$u^* = arg\min E(u) \tag{8}$$

Os próximos passos são puro algebrismo, vamos mexer na função $E(u)$ para desvendar como otimizar essa função. Considere que temos n pontos $x_1, x_2, ..., x_n$ e e_i é o vetor erro do ponto x_i.

$$\begin{aligned} & u^* = arg\min E(u) \\ \iff & \quad \{ \text{ definição (7) de } E(u) \ \} \\ & u^* = arg\min \left(\|e_1\|^2 + \|e_2\|^2 + ... + \|e_n\|^2\right) \\ \iff & \quad \{ \text{ definição de } \|e_i\|^2 \ \} \\ & u^* = arg\min \left(\left(\|x_1\|^2 - \left(\frac{x_1^T u}{\|u\|}\right)^2\right) + \left(\|x_2\|^2 - \left(\frac{x_2^T u}{\|u\|}\right)^2\right) + ... + \left(\|x_n\|^2 - \left(\frac{x_n^T u}{\|u\|}\right)^2\right)\right) \end{aligned}$$

Repare que como fixamos os n pontos $x_1, x_2, ..., x_n$ bem no começo, a norma deles não é afetada pelo u que escolhemos. Em outras palavras, as parcelas $\|x_i\|^2$ **não** dependem de u. Como queremos o **argumento** mínimo de $E(u)$, podemos cortar essas parcelas.

$$\begin{aligned} & u^* = arg\min \left(\left(\|x_1\|^2 - \left(\frac{x_1^T u}{\|u\|}\right)^2\right) + \left(\|x_2\|^2 - \left(\frac{x_2^T u}{\|u\|}\right)^2\right) + ... + \left(\|x_n\|^2 - \left(\frac{x_n^T u}{\|u\|}\right)^2\right)\right) \\ \iff & \quad \{ \text{ parcelas } \|x_i\|^2 \text{ não influenciam em nada } \} \\ & u^* = arg\min \left(-\left(\frac{x_1^T u}{\|u\|}\right)^2 - \left(\frac{x_2^T u}{\|u\|}\right)^2 - .... - \left(\frac{x_n^T u}{\|u\|}\right)^2\right) \\ \iff & \quad \{ \text{ norma ao quadrado para fora } \} \\ & u^* = arg\min \left(-\frac{1}{\|u\|^2}\left(x_1^T u\right)^2 - \frac{1}{\|u\|^2}\left(x_2^T u\right)^2 - .... - \frac{1}{\|u\|^2}\left(x_n^T u\right)^2\right) \\ \iff & \quad \{ \text{ distribuitividade do } - \text{ e da norma ao quadrado } \} \\ & u^* = arg\min -\frac{1}{\|u\|^2}\left(\left(\frac{x_1^T u}{\|u\|}\right)^2 + \left(\frac{x_2^T u}{\|u\|}\right)^2 + ... + \left(\frac{x_n^T u}{\|u\|}\right)^2\right) \end{aligned}$$

O **argumento mínimo** de uma função f é um valor x^* que faz com que $f(x^*)$ seja menor ou igual a qualquer outro valor $f(x)$.

$$x^* \text{ é mínimo se } f(x^*) \leq f(x) \text{ para qualquer } x \tag{9}$$

A definição de **argumento máximo** é parecida, mas agora $f(x^*)$ é maior ou igual a qualquer $f(x)$. Repare que multiplicando tudo por -1, obtemos

$$-f(x^*) \geq -f(x) \text{ para qualquer } x \tag{10}$$

Ou seja, x^* é o **argumento máximo** de $-f$! Vamos sumir com o $-$ e transformar nosso problema em uma maximização.

$$\begin{aligned} & u^* = arg\min -\tfrac{1}{\|u\|^2}\left(\left(x_1^T u\right)^2 + \left(x_2^T u\right)^2 + ... + \left(x_n^T u\right)^2\right) \\ \iff & \quad \{ \text{ minimizar } -f(x) \text{ é igual a maximizar } f(x) \ \} \\ & u^* = arg\max \tfrac{1}{\|u\|^2}\left(\left(x_1^T u\right)^2 + \left(x_2^T u\right)^2 + ... + \left(x_n^T u\right)^2\right) \end{aligned}$$

Repare que as parcelas da soma se assemelham muito a algo que já conhecemos. Se temos um vetor v, sua norma ao quadrado é a soma de suas coordenadas elevadas ao quadrado. Ou seja

$$\|v\|^2 = v_1^2 + v_2^2 + ... + v_n^2 \tag{11}$$

Outro forma de escrever $\|v\|^2$ é como o produto interno do vetor com ele próprio.

$$\|v\|^2 = v^T v \tag{12}$$

Esse padrão é exatamente o que estamos observando na soma da maximização. Temos um vetor

$$\begin{bmatrix} x_1^T u \\ x_2^T u \\ \vdots \\ x_n^T u \end{bmatrix} \tag{13}$$

cuja norma ao quadrado dá exatamente o que temos na maximização. Podemos ir mais além. Podemos definir uma matriz X contendo os pontos $x_1, x_2, ..., x_n$ em suas linhas

$$X = \begin{bmatrix} - & x_1^T & - \\ - & x_2^T & - \\ & \vdots & \\ - & x_n^T & - \end{bmatrix} \tag{14}$$

Com isso podemos escrever o vetor de produtos internos como Xu.

$$\begin{aligned} & \begin{bmatrix} x_1^T u \\ x_2^T u \\ \vdots \\ x_n^T u \end{bmatrix} \\ \iff & \quad \{ \text{ expansão em multiplicação matriz-vetor } \} \\ & \begin{bmatrix} - & x_1^T & - \\ - & x_2^T & - \\ & \vdots & \\ - & x_n^T & - \end{bmatrix} u \\ \iff & \quad \{ \text{ forma matricial } \} \\ & Xu \end{aligned}$$

Então estamos maximizando $\|Xu\|^2$. Fizemos muito algebrismo e é importante que os passos até aqui tenham feito sentido. Pare por um momento e verifique se entendeu cada passo.

Agora podemos seguir manipulando nossa equação.

$$\begin{aligned} & u^* = arg\max \tfrac{1}{\|u\|^2}\left(\left(x_1^T u\right)^2 + \left(x_2^T u\right)^2 + ... + \left(x_n^T u\right)^2\right) \\ \iff & \quad \{ \text{ equivalente a norma de um vetor ao quadrado } \} \end{aligned}$$

$$
\begin{aligned}
& u^* = arg\max \tfrac{1}{\|u\|^2} \left\| \begin{bmatrix} x_1^T u \\ x_2^T u \\ \vdots \\ x_n^T u \end{bmatrix} \right\|^2 \\
\iff & \quad \{ \text{ forma matricial } Xu \ \} \\
& u^* = arg\max \tfrac{1}{\|u\|^2} \|Xu\|^2 \\
\iff & \quad \{ \text{ reescrevendo norma ao quadrado como o produto interno do vetor por si próprio } \} \\
& u^* = arg\max \tfrac{(Xu)^T (Xu)}{u^T u} \\
\iff & \quad \{ \text{ transposta do produto } AB \text{ é igual a } B^T A^T \ \} \\
& u^* = arg\max \tfrac{u^T X^T X u}{u^T u}
\end{aligned}
$$

E essa é a forma final! Mas como todo esse algebrismo ajuda em algo? Bom, para maximizar precisamos antes encontrar os pontos críticos da função e os obtemos derivando ela e igualando a zero. Se chamarmos a forma final do que queremos maximizar de $f(u)$

$$f(u) = \frac{u^T X^T X u}{u^T u} \tag{15}$$

podemos achar $f'(u)$ mais facilmente. A regra do quociente diz que se temos uma função da forma

$$f(x) = \frac{g(x)}{h(x)} \tag{16}$$

Sua derivada será

$$f'(x) = \frac{g'(x)h(x) - g(x)h'(x)}{(h(x))^2} \tag{17}$$

Tendo $f(u)$ da forma que conseguimos podemos nomear $g(u) = u^T X^T X u$ e $h(u) = u^T u$ e assim achar $f'(u)$ usando a regra do quociente! Mas antes, precisamos achar $g'(u)$ e $h'(u)$. Para facilitar as contas, a partir de agora vamos pegar o caso em $\mathbb{R}^2$ onde $u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$.

3.1 Encontrando $h'(u)$

Vamos começar por $h'(u)$ por este ser mais fácil. Podemos expandir $u^T u$ usando sua definição.

$$h(u) = u^T u = u_1^2 + u_2^2 \tag{18}$$

h na verdade é uma função de várias variáveis e justamente por esse motivo h tem mais de uma derivada possível, uma para cada u_i. Se chamarmos de h_i a derivada (parcial) de h em relação a u_i, $h'(u)$ será

$$h'(u) = \begin{bmatrix} h_1(u) \\ h_2(u) \end{bmatrix} \tag{19}$$

O vetor com todas as derivadas. Derivar h_i é derivar em função de u_i e tratar todas as outras varíaveis como números. Como a derivada de um número é 0, só sobram as parcelas com u_i. Derivando h em relação a u_1 e u_2 temos

$$
\begin{aligned}
h_1(u) &= (u_1^2 + u_2^2)_1 = 2u_1 + 0 = 2u_1 \\
h_2(u) &= (u_1^2 + u_2^2)_2 = 0 + 2u_2 = 2u_2
\end{aligned} \tag{20}
$$

Então $h'(u)$ fica

$$h'(u) = \begin{bmatrix} h_1(u) \\ h_2(u) \end{bmatrix} = \begin{bmatrix} 2u_1 \\ 2u_2 \end{bmatrix} = 2 \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = 2u \tag{21}$$

3.2 Encontrando $g'(u)$

Para facilitar as contas, vamos chamar X^TX de Y. Uma propriedade legal de Y é o fato de ser simétrica, ou seja, $Y^T = Y$. Vamos usar essa propriedade posteriormente.

$$(Y)^T = (X^TX)^T \stackrel{*}{=} X^T(X^T)^T = X^TX = Y \tag{22}$$

Na igualdade com um asterisco $(AB)^T = B^TA^T$ foi usado. Expandindo $g(u)$ temos

$$\begin{aligned}
& g(u) \\
= & \quad \{ \text{ definição de } g \ \} \\
& u^TYu \\
= & \quad \{ \text{ expansão de } u \text{ e } Y \ \} \\
& \begin{bmatrix} u_1 & u_2 \end{bmatrix} \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \\
= & \quad \{ \text{ produto } Yu \ \} \\
& \begin{bmatrix} u_1 & u_2 \end{bmatrix} \begin{bmatrix} y_{11}u_1 + y_{12}u_2 \\ y_{21}u_1 + y_{22}u_2 \end{bmatrix} \\
= & \quad \{ \text{ produto interno } \} \\
& u_1(y_{11}u_1 + y_{12}u_2) + u_2(y_{21}u_1 + y_{22}u_2) \\
= & \quad \{ \text{ distributividade de } u_1 \text{ e } u_2 \ \} \\
& y_{11}u_1^2 + y_{12}u_1u_2 + y_{21}u_1u_2 + y_{22}u_2^2
\end{aligned}$$

Então novamente temos uma função de mais de uma variável. Igual fizemos com h, $g'(u)$ será o vetor com a derivada parcial em relação a cada uma de suas variáveis. Suas derivadas parciais são

$$\begin{aligned}
g_1(u) &= (y_{11}u_1^2 + y_{12}u_1u_2 + y_{21}u_1u_2 + y_{22}u_2^2)_1 = 2y_{11}u_1 + y_{12}u_2 + y_{21}u_2 \\
g_2(u) &= (y_{11}u_1^2 + y_{12}u_1u_2 + y_{21}u_1u_2 + y_{22}u_2^2)_2 = 2y_{22}u_2 + y_{12}u_1 + y_{21}u_1
\end{aligned} \tag{23}$$

Pelo fato de $Y = Y^T$, $y_{12} = y_{21}$

$$\begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix}^T = \begin{bmatrix} y_{11} & y_{21} \\ y_{12} & y_{22} \end{bmatrix} \tag{24}$$

Vamos usar isso para juntar esses termos

$$\begin{aligned}
& g'(u) \\
= & \quad \{ \text{ definição de } g'(u) \ \} \\
& \begin{bmatrix} g_1(u) \\ g_2(u) \end{bmatrix} \\
= & \quad \{ \text{ valores de } g_1(u) \text{ e } g_2(u) \ \} \\
& \begin{bmatrix} 2y_{11}u_1 + y_{12}u_2 + y_{21}u_2 \\ 2y_{22}u_2 + y_{12}u_1 + y_{21}u_1 \end{bmatrix} \\
= & \quad \{ \ y_{12} = y_{21} \ \} \\
& \begin{bmatrix} 2y_{11}u_1 + 2y_{12}u_2 \\ 2y_{22}u_2 + 2y_{21}u_1 \end{bmatrix}
\end{aligned}$$

Note que a última forma é exatamente $2Yu$!

$$\begin{aligned}
& \begin{bmatrix} 2y_{11}u_1 + 2y_{12}u_2 \\ 2y_{22}u_2 + 2y_{21}u_1 \end{bmatrix} \\
= & \quad \{ \text{ distributividade no 2 } \} \\
& \begin{bmatrix} 2(y_{11}u_1 + y_{12}u_2) \\ 2(y_{22}u_2 + y_{21}u_1) \end{bmatrix} \\
= & \quad \{ \text{ 2 para fora } \}
\end{aligned}$$

$$
\begin{aligned}
& 2\begin{bmatrix} y_{11}u_1 + y_{12}u_2 \\ y_{22}u_2 + y_{21}u_1 \end{bmatrix} \\
= & \quad \{ \text{ produto } Yu \text{ } \} \\
& 2Yu
\end{aligned}
$$

Agora podemos achar $f'(u)$. Antes de seguir novamente pare e veja se entendeu todos os passos até aqui.

3.3 Derivando $f(u)$ e igualando a 0

Usando os valores $h'(u) = 2u$ e $g'(u) = 2Yu$ que achamos e a regra do quociente, conseguimos achar $f'(u)$.

$$
\begin{aligned}
& f'(u) \\
= & \quad \{ \text{ regra do quociente } \} \\
& \tfrac{g'(u)h(u) - g(u)h'(u)}{(h(u))^2} \\
= & \quad \{ \text{ definição de } g'(u),\ h'(u) \text{ e } h(u) \text{ } \} \\
& \tfrac{2Yu(u^Tu) - (u^TYu)2u}{(u^Tu)^2}
\end{aligned}
$$

Como queremos que $f'(u)$ seja 0, isso só ocorre quando $2Yu(u^Tu) - (u^TYu)2u$ é 0. $(u^Tu)^2$ só seria 0 se o vetor u fosse o vetor nulo, mas esse caso não nos interessa, pois queremos projetar os vetores na reta em que u está, mas com u sendo o vetor nulo $x_i^Tu = 0$ para qualquer ponto.

$$
\begin{aligned}
& 2Yu(u^Tu) - 2(u^TYu)u = 0 \\
\iff & \quad \{ \text{ jogando } (u^TYu)2u \text{ para o outro lado } \} \\
& 2Yu(u^Tu) = 2(u^TYu)u \\
\iff & \quad \{ \text{ dividindo ambos os lados por } 2(u^Tu) \text{ } \} \\
& Yu = \left(\tfrac{u^TYu}{u^Tu}\right) u
\end{aligned}
$$

Note que $\frac{u^TYu}{u^Tu}$ é um número.

O que essa última igualdade nos diz é que os pontos críticos de $f(u)$ são os vetores u que quando aplicamos Y conseguimos um múltiplo deles. É exatamente a definição de autovetores e autovalores! x é autovetor de A se

$$Ax = \lambda x \tag{25}$$

é verdade para algum número λ. Não consideramos o vetor nulo autovetor.

Portanto os pontos críticos serão os autovetores de Y com autovalores associados $\lambda = \frac{u^TYu}{u^Tu}$. Repare ainda que $\lambda = \frac{u^TYu}{u^Tu}$ é exatamente $f(u)$. Como queremos o máximo de $f(u)$, precisamos do ponto crítico u cujo valor de $f(u)$ seja o maior. Como, para os pontos críticos, $f(u)$ é exatamente o autovalor de u, a resposta para nosso problema de maximização é o maior autovalor de Y!

Desse modo, o argumento máximo da função é o autovetor associado a esse maior autovalor. O maior autovetor de Y é um vetor contido na reta que minimiza a soma dos erros. É a reta que projetaremos nossos dados para reduzir sua dimensão. Nesse caso chamamos u de **primeiro componente principal**.

4 Reduções para dimensões maiores que 1

O modo como resolvemos o problema anterior força nossos dados a cair para a dimensão 1 sempre. Mas e se quisermos reduzir nossos dados para uma dimensão maior que 1? Basta pegarmos a próxima maior solução. Se escolhemos o maior autovetor de Y, basta procurarmos pelo próximo ponto crítico com maior valor de $f(u)$. Como u é ponto crítico, $f(u)$ é seu autovalor e como já escolhemos o maior autovetor, o próximo maior argumento de $f(u)$ será o segundo maior autovetor de Y! Chamamos esse vetor de **segundo componente principal**. Isso se segue para as próximas

dimensões: Se queremos reduzir nossos dados para um espaço de m dimensões, projetamos os pontos no espaço gerado pelos m maiores autovetores de Y e os nomeamos de os primeiros m **componentes principais** de X.

4.1 Formulação mais matemática

Se quisermos, podemos definir tudo isso melhor. Considere que queremos reduzir nossos dados para dimensão 2 e já temos o maior autovetor de Y. Como Y é uma matriz simétrica, temos pelo teorema espectral que seus autovetores são ortogonais entre si. Ou seja, o produto interno entre dois autovetores distintos é 0. Podemos adicionar isso como restrição em nossa maximização. Se v_1 é o primeiro componente principal, a maximização fica

$$u^* = arg \max_{v_1^T u = 0} \frac{u^T X^T X u}{u^T u} \tag{26}$$

Desse modo, excluímos v_1 como solução dessa maximização. Como todos os pontos críticos de $f(u)$ são autovetores de Y, todos satisfazem, exceto o próprio v_1, $v_1^T u = 0$ e seguem sendo soluções válidas. Como queremos o argumento máximo e $f(u)$ é autovalor de Y nos pontos críticos, pegaremos o autovetor associado ao maior autovalor que satisfaça $v_1^T u = 0$. Ou seja, o segundo maior autovetor. Para o m-ésimo componente principal, fazemos ele ser perpendicular a todos os $m-1$ vetores já obtidos. O resultado é o m-ésimo maior autovetor de Y.

$$u^* = arg \max_{v_1^T u = 0, v_2^T u = 0, ..., v_{m-1}^T u = 0} \frac{u^T X^T X u}{u^T u} \tag{27}$$

5 Projetando nos componentes principais e pseudo-código

Falamos muito, mas não comentamos sobre a projeção de fato. Para projetar um vetor v em uma reta, primeiro pegamos um vetor u na reta e calculamos o comprimento da projeção, que é $\frac{1}{\|u\|} v^T u$ como vimos anteriormente. Em seguida multiplicamos esse comprimento pelo vetor u normalizado. A ideia é que estamos esticando esse vetor até que tenha norma $\frac{1}{\|u\|} v^T u$. A forma final da projeção p de v na reta é

$$\begin{aligned} & p = \left(\tfrac{1}{\|u\|} v^T u\right) \tfrac{1}{\|u\|} u \\ \iff & \quad \{ \text{ álgebra } \} \\ & p = \tfrac{1}{\|u\|^2} \left(v^T u\right) u \\ \iff & \quad \{ \ \|u\|^2 = u^T u \ \} \\ & p = \left(\tfrac{v^T u}{u^T u}\right) u \end{aligned}$$

Agora, como projetamos em mais de um componente principal? Pela fato dos componentes principais serem perpendiculares entre si, projetar no espaço que eles geram é igual a projetar em cada um individualmente e somar esse resultado. Ou seja, se $b_1, b_2, ..., b_m$ são os m componentes principais de X e vamos assumir que suas normas sejam 1 pois assim $b_i^T b_i = 1$, a projeção de um ponto x no espaço gerado por eles é

$$p = (x^T b_1) b_1 + (x^T b_2) b_2 + ... + (x^T b_m) b_m \tag{28}$$

Conseguimos representar isso na forma matriz-vetor? Sim! Vamos colocar os m componentes em uma matriz B, onde cada coluna é um componente principal.

$$B = \begin{bmatrix} | & | & .... & | \\ b_1 & b_2 & .... & b_m \\ | & | & .... & | \end{bmatrix} \tag{29}$$

Podemos ver que $B^T x$ dá um vetor com os valores de $b_i^T x = x^T b_i$ nas entradas.

$$\begin{aligned} & B^T x \\ = & \quad \{ \text{ definição de } B \ \} \end{aligned}$$

$$
\begin{aligned}
&\begin{bmatrix} - & b_1^T & - \\ - & b_2^T & - \\ & \vdots & \\ - & b_n^T & - \end{bmatrix} x \\
= &\quad \{ \text{ produto vetor-matriz } \} \\
&\begin{bmatrix} x^T b_1 \\ x^T b_2 \\ \vdots \\ x^T b_m \end{bmatrix}
\end{aligned}
$$

As entradas de $B^T x$ indicam o quanto de cada componente principal o vetor x usa. Para conseguir a soma final, basta multiplicar por B pela esquerda.

$$
\begin{aligned}
&B \begin{bmatrix} x^T b_1 \\ x^T b_2 \\ \vdots \\ x^T b_m \end{bmatrix} \\
= &\quad \{ \text{ definição de } B \ \} \\
&\begin{bmatrix} | & | & .... & | \\ b_1 & b_2 & .... & b_n \\ | & | & .... & | \end{bmatrix} \begin{bmatrix} x^T b_1 \\ x^T b_2 \\ \vdots \\ x^T b_m \end{bmatrix} \\
= &\quad \{ \text{ produto matriz-vetor } \} \\
&(x^T b_1) b_1 + (x^T b_2) b_2 + ... + (x^T b_n) b_n
\end{aligned}
$$

Mas como generalizamos isso para n pontos $x_1, x_2, ..., x_n$? Basta trocar x pela matriz X de pontos. Considere que os pontos $x_1, x_2, ..., x_n$ são as linhas de X como já fizemos anteriormente.

$$
X = \begin{bmatrix} - & x_1^T & - \\ - & x_2^T & - \\ & \vdots & \\ - & x_n^T & - \end{bmatrix} \tag{30}
$$

Temos que $B^T X^T$ contêm, em cada coluna, o quanto de cada um dos m componentes cada um dos n pontos usa.

$$
B^T X^T = \begin{bmatrix} - & b_1^T & - \\ - & b_2^T & - \\ & \vdots & \\ - & b_n^T & - \end{bmatrix} \begin{bmatrix} | & | & .... & | \\ x_1 & x_2 & .... & x_n \\ | & | & .... & | \end{bmatrix} = \begin{bmatrix} x_1^T b_1 & x_2^T b_1 & .... & x_n^T b_1 \\ x_1^T b_2 & x_2^T b_2 & .... & x_n^T b_2 \\ .... & .... & .... & .... \\ x_1^T b_n & x_2^T b_n & .... & x_n^T b_n \end{bmatrix}
$$

Se queremos aproximar a matriz X original usando somente os m componentes, podemos fazer $BB^T X^T$. Se renomearmos $B^T X^T$ como C^T, chegamos na famosa forma

$$
X \approx BC^T \tag{31}
$$

Que é a melhor aproximação de posto m da matriz X. Ou, em outras palavras, o espaço m-dimensional que melhor descreve nossos dados. Nesse caso, B representa a **base** desse espaço m-dimensional, os vetores que usaremos para descrever os pontos, e C^T representa os **coeficientes** dos pontos nessa base, o quanto cada ponto usa de cada dos vetores da base na matriz B.