MT404/MS993

Métodos computacionais de álgebra linear

Segundo Projeto – 2023

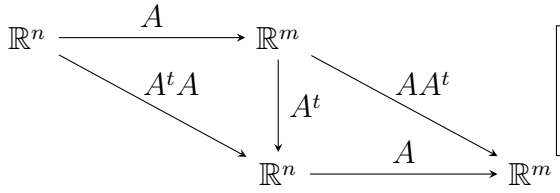
Teorema de Eckart-Young-Mirsky e decomposição em produtos de Kronecker

Resumo

Neste segundo projeto, vamos considerar o problema, atualmente bastante popular, de aproximar uma matriz dada por um produto de Kronecker. Veremos como reduzir este problema à chamada aproximação de baixo posto e como o Teorema de Eckart-Young-Mirsky se aplica neste caso. A ferramenta básica será a ubíqua Decomposição em Valores Singulares (SVD) de uma matriz, que deverá ser implementada **completamente**.

1 Decomposição em Valores Singulares

A Decomposição em Valores Singulares (SVD, do inglês Singular Value Decomposition) será o ponto central desta atividade. Comecemos com uma rápida revisão. Nosso diagrama preferido, agora feito caprichosamente com o magnífico tikz (arquivo tex desta atividade aqui),



$$\dim \mathrm{Ker}(A) + \dim \mathrm{Im}(A) = n$$
$$\mathrm{Ker}(A) \cap \mathrm{Im}(A^t) = \varnothing$$
$$\mathbb{R}^n = \mathrm{Ker}(A) \oplus \mathrm{Im}(A^t)$$

destaca as matrizes $M = AA^t$ e $N = A^tA$. Ambas são simétricas e, portanto, completamente diagonalizáveis,

$$D_N = V^tNV, \quad D_M = U^tNU, \tag{1}$$

sendo U e V matrizes ortogonais, enquanto D_N e D_M são matrizes diagonais contendo, respectivamente, os autovalores (reais) de N e de M. Por construção, vemos que N e M são matrizes não negativa, *i.e.*,

$$X^tNX = X^tA^tAX = (AX)^tAX \geq 0 \tag{2}$$

para todo $X \in \mathbb{R}^n$, e de maneira semelhante para M. Isto implica que todos os autovalores de N e M são não negativos, pois se houvesse, por exemplo, um autovalor negativo para N, teríamos $X^tNX < 0$ para o autovetor associado. É importante notar que se $\mu > 0$ é um autovalor de N, então necessariamente também será de M, e vice-versa, pois se

$$NX = A^tAX = \mu X, \tag{3}$$

basta multiplicarmos por A pela esquerda e teremos

$$AA^tAX = M(AX) = \mu AX, \tag{4}$$

i.e., AX será autovetor de M com o mesmo autovalor μ. Como os autovalores em questão são não negativos, denotaremo-los por $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_n^2$, com $\sigma_k \geq 0$. Notem que N e M em geral não têm as mesmas dimensões, portanto não terão o mesmo número de autovalores nulos, pois de maneira geral terão nulidades diferentes ($\dim \mathrm{Ker}(N) \neq \dim \mathrm{Ker}(M)$). Os números reais não negativos σ_k são os chamados valores singulares da matriz A.

Estes resultados nos permitem afirmar que qualquer matriz $A_{m\times n}$ pode sempre ser decomposta de maneira única como

$$A = U\Sigma V^t, \tag{5}$$

sendo $U_{m\times m}$ e $V_{n\times n}$ as matrizes ortogonais das diagonalizações (1) e $\Sigma_{m\times n}$ uma matriz retangular com os valores singulares σ_k na sua diagonal principal e todas as outras entradas nulas. A unicidade da decomposição vem diretamente dos resultados que garantem a diagonalização (1), basta escrever A^tA e AA^t a partir de (5) e identificar as matrizes envolvidas. Notem que $\Sigma\Sigma^t$ e $\Sigma^t\Sigma$ serão matrizes quadradas com os autovalores $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_n^2 \geq 0$ em suas diagonais.

Assim como fizemos em aula, um exemplo é bem vindo aqui para ilustrar estes conceitos. Vamos considerar a matriz $A_{3\times 2} : \mathbb{R}^2 \to \mathbb{R}^3$ dada por

$$A = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \end{pmatrix}. \tag{6}$$

Teremos neste caso

$$N = A^t A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}, \quad M = AA^t = \begin{pmatrix} 2 & 0 & 2 \\ 0 & 2 & 0 \\ 2 & 0 & 2 \end{pmatrix}. \tag{7}$$

A matriz M é diagonalizada como

$$M = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{-1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 4 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{-1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{pmatrix}^t, \tag{8}$$

e N como

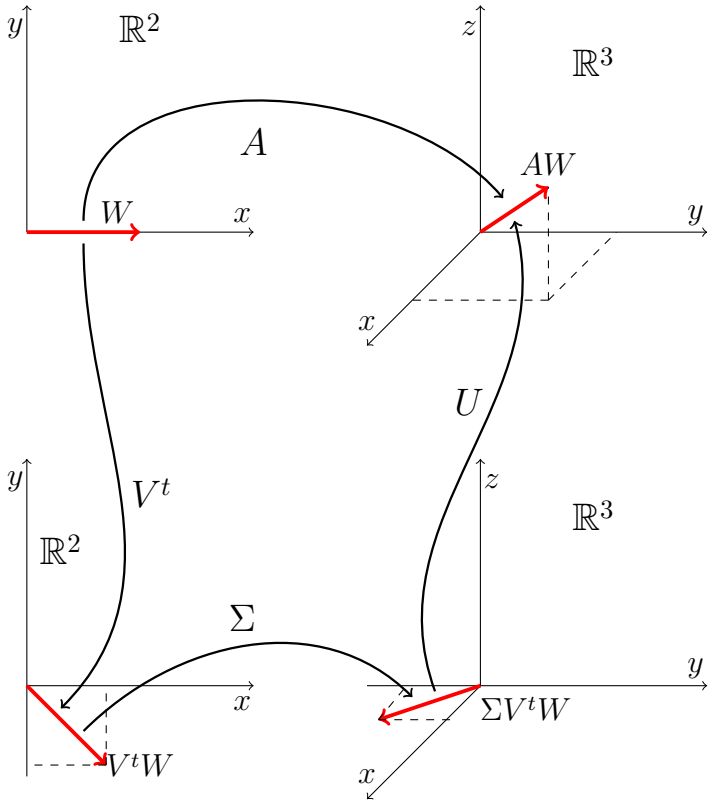
$$N = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}^t. \tag{9}$$

Os valores singulares de A são 2 e $\sqrt{2}$ e sua decomposição SVD será

$$\underbrace{\begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \end{pmatrix}}_{A_{3\times 2}} = \underbrace{\begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{-1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{pmatrix}}_{U_{3\times 3}} \underbrace{\begin{pmatrix} 2 & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \end{pmatrix}}_{\Sigma_{3\times 2}} \underbrace{\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}^t}_{(V_{2\times 2})^t}. \tag{10}$$

Este exemplo é útil para ilustrar o papel das duas rotações U e V e da matriz Σ. Vamos analisar os três passos da decomposição SVD na ação da matriz A sobre um vetor específico, no caso $W = (1, 0)^t \in \mathbb{R}^2$. Primeiro, notem que a matriz A leva o vetor W no vetor $AW = (1, 1, 1)^t \in \mathbb{R}^3$, acompanhem os passos no diagrama a seguir, também orgulhosamente feito com ajuda do tikz. Examinemos a ação dos três produtos $U\Sigma V^t W$ separadamente. O primeiro, produto, $V^t W$, é uma rotação no domínio ($\mathbb{R}^2$) de A, teremos $V^t W = \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right)^t$. O próximo passo é a ação da matriz Σ sobre esse produto. É interessante notar que é essa matriz que "transfere" o vetor de $\mathbb{R}^2$ para

$\mathbb{R}^3$, ou seja, do domínio para a imagem de A, neste caso deformando suas dimensões, ao contrário das rotações. Temos aqui $\Sigma V^t W = \left(\sqrt{2}, -1, 0\right)^t$. Notem que a "transferência" é feita de maneira a "inserir" o vetor de $\mathbb{R}^2$ no plano $z = 0$ de $\mathbb{R}^3$, eventualmente deformando-o. Finalmente, temos a ação de U, que corresponde a uma rotação na imagem de A, $U\Sigma V^t W = (1, 1, 1)^t$. Este exemplo simples ilustra como a ação de uma matriz genérica $A_{m\times n}$ se resume geometricamente em: uma rotação em seu domínio, uma "transferência" de seu domínio a sua imagem, e uma última rotação em sua imagem.



1.1 Aproximação de posto baixo

A Decomposição em Valores Singulares de uma matriz tem inúmeras aplicações relevantes, mas aqui focaremos no chamado problema da melhor aproximação de posto baixo. Para introduzir estes problemas, convêm relembrar uma propriedade elementar do produto matricial. Sejam $U_{m\times p}$ e $V_{p\times n}$ duas

matrizes e $u_k \in \mathbb{R}^m$ e $v_k \in \mathbb{R}^n$, $1 \leq k \leq q$ os seguintes vetores

$$u_i = \mathrm{Col}_k(U), \quad v_k = \mathrm{Col}_k(V^t), \tag{11}$$

i.e., os vetores correspondentes as colunas de U e as linhas de V. Pode-se expressar o produto matricial UV a partir dos produtos desses vetores como

$$UV = \sum_{k=1}^{p} u_k v_k^t. \tag{12}$$

O produto $u_k v_k^t$, que é uma matriz $m \times n$, também é chamado de produto externo (outer) ou também de produto tensorial dos vetores u_k e v_k. É também comum encontrarmos a notação $u_k \otimes v_k$, ou as vezes $u_k \otimes_o v_k$, para diferenciá-lo do produto de Kronecker, que veremos a seguir. A decomposição SVD (5) pode sempre ser vista como

$$A = \sum_k \sigma_k u_k v_k^t, \tag{13}$$

sendo u_k as colunas de U e, portanto, os autovetores de AA^t. De maneira semelhante, v_k são os autovetores de A^tA. A soma em k deve ser entendida como sendo sobre todos os valores singulares de A, mas é evidente que apenas os não nulos irão contribuir. É claro também que o número de valores singulares não nulos é o posto de A.

Em palavras simples, o problema da aproximação de posto baixo (*low-rank approximation*) pode ser formulado como: dada uma matriz arbitrária A de posto p, que matriz de posto $q < p$ pode ser considerada sua melhor aproximante? Com um pouco mais de precisão, qual é a matriz B de posto mais baixo tal que $||A - B||$ seja mínimo? Usualmente, nestes problemas emprega-se a norma 2, a norma espectral, segundo a qual

$$||A||_2 = \sigma_1. \tag{14}$$

A resposta a este problema é dado pelo teorema de Eckart-Young-Mirsky, que tem uma história curiosa facilmente encontrada em materiais on-line. O teorema foi formulado inicialmente em espaços de Hilbert de dimensão infinita por Erhard Schmidt na primeira década do século XX, adiantando diversas ideias úteis na Mecânica Quântica. Cerca de trinta anos depois, sua versão de dimensão finita (a que nos interessa aqui) foi redescoberta por Eckart e Young em problemas de psicometria. Nos anos 60, Mirsky refinou e

generalizou os resultados de Eckart e Young, e os três passaram a dar nome ao teorema de Schmidt, numa clara aplicação da chamada lei de Stigler[1]. Como prova inequívoca da relevância deste teorema, ele tem sido "redescoberto" recentemente, múltiplas vezes de fato, no contexto de Inteligência Artificial e Aprendizado de Máquinas. Mais detalhes sobre a história deste teorema e alguns de seus precedentes, ver [1].

O teorema nos garante que o melhor aproximante de posto $q < p$ para a matriz A de posto p será a matriz

$$A_q = \sum_{k=1}^{q} \sigma_k u_k v_k^t, \tag{15}$$

i.e., o "truncamento" da SVD de A em seu q-ésimo valor singular. Seu enunciado mais simples é o seguinte.

Teorema 1 (Eckart-Young-Mirsky) *Seja $A_{m\times n}$ uma matriz de posto p. Para a norma espectral (idem para a de Frobenius),*

$$||A - A_q||_2 \leq ||A - B||_2$$

para todas matrizes B de posto $q < p$.

É evidente que há aqui uma questão de unicidade que não pode ser garantida se A possuir valores singulares repetidos, vamos simplesmente ignorar esta questão. Vamos nos restringir a encontrar **uma** solução ótima para o problema, mas não é difícil encontrar todas. A prova é relativamente simples. Seja $x \in \mathrm{Ker}(B)$ um vetor unitário. Sabemos que $\dim \mathrm{Ker}(B) = n - q$. Tomemos agora o espaço $\mathrm{span}(v_1, v_2, \ldots, v_{q+1})$, *i.e.*, o espaço gerado pelas $q+1$ primeiras colunas de V da SVD de A. Notem que estes dois subespaços lineares tem intersecção não vazia, pois ambos são subespaços de $\mathbb{R}^n$ e suas dimensões somadas excedem n. Portanto, é garantido que existirá sempre um vetor unitário $x \in \mathrm{Ker}(B) \cap \mathrm{span}(v_1, v_2, \ldots, v_{q+1})$. Este vetor será da forma

$$x = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_{q+1} v_{q+1}, \tag{16}$$

[1]De acordo com a wikipedia em português: a Lei de Stigler é um fenômeno proposto por Stephen Stigler, professor de estatística da Universidade de Chicago em 1980. Em sua forma mais simples e direta ela diz: "Nenhuma descoberta científica recebe o nome de seu descobridor". Para exemplificar esse fato, Stigler nomeou o sociologista Robert K. Merton como o descobridor da Lei de Stigler.

com $\alpha_1^2 + \alpha_2^2 + \cdots + \alpha_{q+1}^2 = 1$. Teremos:

$$\begin{aligned} \|A-B\|_2^2 &\geq \|(A-B)x\|_2^2 \quad \text{da definição da norma 2,} \\ &= \|Ax\|_2^2 \qquad\quad \text{pois } x \in \mathrm{Ker}(B), \\ &= x^t V \Sigma^2 V^t x \qquad \text{da SVD de} A, \\ &= \alpha_1^2\sigma_1^2 + \alpha_2^2\sigma_2^2 + \cdots + \alpha_{q+1}^2\sigma_{q+1}^2 \\ &\geq \sigma_{q+1}^2(\alpha_1^2 + \alpha_2^2 + \cdots + \alpha_{q+1}^2) = \sigma_{q+1}^2 = \|A - A_q\|_2^2. \end{aligned}$$

A restrição inicial de escolher $x \in \mathrm{span}(v_1, v_2, \ldots, v_{q+1})$ não é de fato necessária. Como as colunas $\{v_i\}$ de V são uma base ortogonal de $\mathbb{R}^n$, um vetor $x \in \mathrm{Ker}(B)$ unitário arbitrário terá a forma

$$x = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n, \tag{17}$$

com $\alpha_1^2 + \alpha_2^2 + \cdots + \alpha_n^2 = 1$. Repetindo-se os passos acima, teremos

$$\begin{aligned} \|A-B\|_2^2 &\geq \|(A-B)x\|_2^2 \quad \text{da definição da norma 2,} \\ &= \|Ax\|_2^2 \qquad\quad \text{pois } x \in \mathrm{Ker}(B), \\ &= x^t V \Sigma^2 V^t x \qquad \text{da SVD de} A, \\ &= \alpha_1^2\sigma_1^2 + \alpha_2^2\sigma_2^2 + \cdots + \alpha_n^2\sigma_n^2 \\ &\geq \sigma_{q+1}^2(\underbrace{\alpha_1^2 + \alpha_2^2 + \cdots + \alpha_{q+1}^2}_{\mu^2}) + \sigma_{q+2}^2(\underbrace{\alpha_{q+2}^2 + \cdots + \alpha_n^2}_{\nu^2}), \end{aligned}$$

com $\mu^2 + \nu^2 = 1$. O ponto fundamental da prova é que sempre existe um $x \in \mathrm{Ker}(B)$ unitário tal que $\nu = 0$, quer dizer, sempre existirá um $x \in \mathrm{Ker}(B)$ tal que

$$\|A-B\|_2^2 \geq \|(A-B)x\|_2^2 \geq \sigma_{q+1}^2 = \|A - A_q\|_2^2,$$

pois $\dim \mathrm{Ker}(B) = n-q$ e $\dim \mathrm{span}(v_{q+2}, \ldots, v_n) = n-q-1$ e, portanto, sempre é possível encontrar um $x \in \mathrm{Ker}(B)$ que seja ortogonal a $\mathrm{span}(v_{q+2}, \ldots, v_n)$, quer dizer, que more em $\mathrm{span}(v_1, v_2, \ldots, v_{q+1})$. Explicitamente, temos que, se $\{b_i\}$ é uma base de $\mathrm{Ker}(B)$, o vetor x pode ser escrito como

$$x = \beta_1 b_1 + \beta_2 b_2 + \cdots + \beta_{n-q} b_{n-q}. \tag{18}$$

Agora, devemos exigir $v_k \cdot x = 0$, para $k = q+2, q+3, \ldots, n$, *i.e.*, ficaremos com o sistema linear

$$\begin{pmatrix} b_1 \cdot v_{q+2} & b_2 \cdot v_{q+2} & \cdots & b_{n-q} \cdot v_{q+2} \\ b_1 \cdot v_{q+3} & b_2 \cdot v_{q+3} & \cdots & b_{n-q} \cdot v_{q+3} \\ \vdots & \vdots & \ddots & \vdots \\ b_1 \cdot v_n & b_2 \cdot v_n & \cdots & b_{n-q} \cdot v_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{n-q} \end{pmatrix} = 0, \tag{19}$$

que sempre admite solução não trivial.

2 Produto de Kronecker

Lidaremos neste projeto com o chamado produto de Kronecker de duas matrizes. Por simplicidade, vamos considerar apenas matrizes reais. Dadas duas matrizes $A \in \mathbb{R}^{n\times m}$ e $B \in \mathbb{R}^{p\times q}$, seu produto de Kronecker $A \otimes B \in \mathbb{R}^{(np)\times(mq)}$ é definido como

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1m}B \\ a_{21}B & a_{22}B & \cdots & a_{2m}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}B & a_{n2}B & \cdots & a_{nm}B \end{pmatrix}. \tag{20}$$

É comum também usar o símbolo $\otimes_k$ para este produto de matrizes. O produto de Kronecker é o produto tensorial de matrizes e satisfaz uma série de propriedades elementares, ver [2], a referencial central para as motivações deste trabalho, e [3]. Uma das propriedades interessantes deste produto é a do posto: $\text{posto}\,(A \otimes B) = \text{posto}\,(A)\,\text{posto}\,(B)$. Suas aplicações são as mais diversas, ver [2]. Recentemente, em particular, vem ganhando relevância em problemas de informação quântica. Para sistemas de duas partes, A e B, sabe-se que os estados que **não** exibem emaranhamento são os descritos precisamente por expressões do tipo $A \otimes B$ para as grandezas de interesse dessas partes.

Há também grande interesse em se definir "medidas" de emaranhamento. Seja C um estado quântico de um sistema composto por duas partes iguais A e B. Estes estados são completamente descritos por matrizes quadradas de mesmo nome. Se as partes são caracterizados por matrizes $n \times n$, o sistema completo estará associado a uma matriz $n^2 \times n^2$. Uma possível "medida" de emaranhamento de um estado C seria

$$\mu(C) = \min \|C - A \otimes B\|_2, \tag{21}$$

para todas matrizes A e B. Em particular, temos que se C for um produto de Kronecker, teremos $\mu(C) = 0$, conforme a intuição da Mecânica Quântica que não vem ao caso aqui. Podemos encarar (21) como um problema de encontrar a melhor aproximação de C em termos de um produto de Kronecker $A \otimes B$. É evidente que este problema não tem solução única, pois as matrizes νA

e $\nu^{-1}B$ tem o mesmo produto de Kronecker para qualquer $\nu \neq 0$. Esta não unicidade, obviamente, não é um problema para a definição da medida de emaranhamento (21). Sempre que falarmos em melhor aproximação por um produto de Kronecker, estamos sempre admitindo que, de fato, trata-se de uma família de aproximantes do tipo νA e $\nu^{-1}B$ com $\nu \neq 0$.

Vamos mostrar como o teorema de Eckart-Young-Mirsky pode ser usado para resolver o problema da melhor aproximação em termos de um produto de Kronecker. Vamos supor por simplicidade que temos uma matriz $C_{6\times4}$ e queremos determinar qual é sua melhor aproximação em termos de um produto $A_{3\times2} \otimes B_{2\times2}$, com todos os *caveats* com relação à unicidade deste problema. Deve-se minimizar

$$\left\| \left(\begin{array}{cc|cc} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} & c_{24} \\ \hline c_{31} & c_{32} & c_{33} & c_{34} \\ c_{41} & c_{42} & c_{43} & c_{44} \\ \hline c_{51} & c_{52} & c_{53} & c_{54} \\ c_{61} & c_{62} & c_{63} & c_{64} \end{array}\right) - \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} \otimes \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \right\|_2 , \tag{22}$$

dada a matriz C. Uma operação muito comum em álgebra linear computacional é o *flattening*, as vezes também chamada de "vetorização", que corresponde a reorganizar uma matriz em uma única coluna. Por exemplo, para o caso de uma matriz $A_{3\times2}$, define-se uma matriz coluna

$$\mathrm{flat}(A) = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \\ a_{12} \\ a_{22} \\ a_{32} \end{pmatrix} \in \mathbb{R}^6. \tag{23}$$

Inspecionando-se (22), vemos que a matriz C está naturalmente dividida em 6 blocos $\bar{C}_{ij}$ e estamos procurando aproximações que minimizem

$$\left\| \bar{C}_{ij} - a_{ij} B \right\|_2 , \tag{24}$$

para todo i, j, o que é equivalente a minimizar

$$\left\| \mathrm{flat}(\bar{C}_{ij}) - a_{ij} \mathrm{flat}(B) \right\|_2 . \tag{25}$$

Vamos agora introduzir uma nova matriz $\tilde{C}_{6\times 4}$

$$\tilde{C} = \big(\text{flat}(\bar{C}_{11}), \text{flat}(\bar{C}_{21}), \text{flat}(\bar{C}_{31}), \text{flat}(\bar{C}_{12}), \text{flat}(\bar{C}_{22}), \text{flat}(\bar{C}_{32})\big)^t. \quad (26)$$

Em termos dessa nova matriz, a minimização (22) fica

$$\|\tilde{C} - \underbrace{\text{flat}(A)\text{flat}(B)^t}_{\text{posto 1}}\|_2, \quad (27)$$

de onde vemos que o problema se reduz a uma aproximação de posto baixo e, portanto, a uma aplicação do teorema de Eckart-Young-Mirsky. A solução "padrão" corresponde a $\text{flat}(A) = \sqrt{\sigma_1}u_1$ e $\text{flat}(B) = \sqrt{\sigma_1}v_1$, ver (15), e as matrizes A e B são recuperadas desfazendo-se a vetorização.

3 As atividades

Como todos os nossos projetos, reproduzir todas as passagens relevantes deste texto é a primeira parte das atividades. Porém, não há necessidade de entregar estas passagens, mas eu realmente recomendo que as façam. A principal tarefa será propor e implementar um algoritmo para calcular $\mu(C)$ definida por (22). Estamos interessados apenas em casos não triviais, então vamos excluir os casos nos quais pelo menos uma das matrizes "fatores" A e B é um escalar. Portanto, dado uma matriz $C_{m\times n}$, vamos minimizar (22) para todos os possíveis produtos $A_{pq} \otimes B_{rs}$ com $m = pr$ e $n = qs$, excluindo-se o caso trivial $p = q = 1$ e/ou $r = s = 1$.

Como o problema se reduz basicamente a uma questão de aproximação de posto baixo, será necessário fazer a decomposição em valores singulares de várias matrizes. Acho que vale a pena que a SVD seja completamente implementada também. Como comentei em aula, os algoritmos modernos de diagonalização, que são a base da SVD, são bastante sofisticados tecnicamente e extremamente eficientes. Não há nenhuma necessidade em implementar seus próprios algoritmos, que serão muito provavelmente menos precisos e com certeza absoluta menos eficientes. Em situações nas quais eficiência e precisão são essenciais, a melhor opção é sempre apelar para os *"Eigenpackages of Canned Eigenroutines"*, ver por exemplo [5]. Porém, acho instrutivo que implementem um algoritmos desses pelo menos uma vez, para que tenham ideia das dificuldades envolvidas. Além disso, os algoritmos que serão propostos podem ser facilmente refinados e atingirão um desempenho bastante aceitável quando comparado aos pacotes especializados.

Vamos optar pela abordagem mais direta. Vamos determinar a SVD (5) a partir da diagonalização das matrizes simétricas AA^t e A^tA. A seguir, vamos expor brevemente e sem muitas preocupações com rigor os passos necessários para a diagonalização de matrizes simétricas.

3.1 Diagonalização de matrizes simétricas

Se A é uma matriz simétrica, sabe-se que ela sempre pode ser diagonalizada $D = U^tAU$, sendo U a matriz ortogonal formada por seus autovetores e D a matriz diagonal contendo os respectivos autovalores. Vamos explorar o chamado algoritmo QR iterativo para obter as matrizes U e D a partir de uma matriz A simétrica dada. O primeiro passo é apresentar a chamada decomposição QR.

3.1.1 Decomposição QR

Dada uma matriz $A_{m\times n}$, podemos sempre decompô-la em um produto $A = QR$, sendo $Q_{m\times m}$ uma matriz ortogonal e $R_{m\times n}$ uma matriz triangular superior (*Right*), *i.e.*, uma matriz com todas as entradas abaixo da diagonal principal nulas. Antes de discutir como fazer esta decomposição, convêm explorar seu significado. Considerem as colunas de A e as de Q

$$a_k = \mathrm{Col}_k(A) \in \mathbb{R}^m, \quad q_\ell = \mathrm{Col}_\ell(Q) \in \mathbb{R}^m \tag{28}$$

$1 \le k \le n$ e $1 \le \ell \le m$. Como Q é ortogonal, os vetores $\{q_\ell\}$ são uma base ortonormal de $\mathbb{R}^m$. A decomposição QR implica

$$\begin{array}{rcl} a_1 &=& r_{11}q_1 \\ a_2 &=& r_{12}q_1 + r_{22}q_2 \\ a_3 &=& r_{13}q_1 + r_{23}q_2 + r_{33}q_3 \\ \vdots & \vdots & \qquad\qquad \vdots \end{array} \tag{29}$$

É claro que a decomposição QR proporciona uma base ortonormal para o espaço $\mathrm{Span}(a_k)$ "adaptada", pois q_1 tem a direção de a_1, $\mathrm{Span}(q_1, q_2) = \mathrm{Span}(a_1, a_2)$, etc, de maneira semelhante à construção de Gram-Schmidt.

A maneira mais direta de se implementar a decomposição QR explora as chamadas transformações de Householder, que são essencialmente reflexões

por certos (hiper)planos. Relembrando, a operação elementar de reflexão de um vetor $X \in \mathbb{R}^m$ por um plano com vetor normal unitário $N \in \mathbb{R}^m$ é o vetor

$$X_R = PX = (1 - 2NN^t)X. \tag{30}$$

Por se tratar de uma reflexão, P é obviamente ortogonal, pois

$$PP^t = P^2 = (1 - 2NN^t)(1 - 2NN^t) = 1 - 4NN^t + 4NN^tNN^t = 1. \tag{31}$$

Uma transformação de Householder elementar é aquela que transforma um vetor $X = (x_1, x_2, \ldots, x_m)^t \in \mathbb{R}^m$ arbitrário em um vetor do tipo $X_R = (||X||, 0, \ldots, 0)^t \in \mathbb{R}^m$. É fácil determinar que vetor N tem essa ação. Seja $M = (x_1 - ||X||, x_2, \ldots, x_m)^t$ e

$$\begin{aligned} N = \frac{M}{||M||} &= \frac{1}{\sqrt{2||X||(||X|| - x_1)}}(x_1 - ||X||, x_2, \ldots, x_m)^t \\ &= \frac{1}{\sqrt{2||X||}}\left(-\sqrt{||X|| - x_1}, \frac{x_2}{\sqrt{||X|| - x_1}}, \ldots, \frac{x_m}{\sqrt{||X|| - x_1}}\right)^t . \end{aligned} \tag{32}$$

Notem que $N^tX = \frac{\sqrt{||X||(||X|| - x_1)}}{\sqrt{2}}$ e, portanto, $NN^tX = \frac{1}{2}\left(X - (||X||, 0, \ldots, 0)\right)^t$, de onde segue finalmente[2] $PX = (||X||, 0, \ldots, 0)^t$.

Assim, temos claramente um algoritmo para se determinar a decomposição Q de uma matriz A. Começa-se aplicando em A a transformação de Householder P_1 que zera todas as posições abaixo da diagonal na primeira coluna. O passo seguinte é aplicar uma transformação P_2 que zera as posições abaixo da diagonal na segunda coluna de P_1A. Ao final destas operações, teremos algo como

$$P_{n-1}P_{n-2}\cdots P_2P_1A = R, \tag{33}$$

de onde temos finalmente a matriz ortogonal $Q = P_1P_2\cdots P_{n-2}P_{n-1}$ da decomposição QR.

3.1.2 Algoritmo QR iterativo

A decomposição QR é a base dos algoritmos mais eficientes para a diagonalização de matrizes. Vamos apresentar rapidamente a seguir o chamado algoritmo QR iterativo e explicar como funciona. O algoritmo 1 é extre-

[2]Há uma questão interessante aqui. Pode-se interpretar a transformação P como uma rotação? Pensem...

Algoritmo 1: Algoritmo QR para diagonalização de matrizes

Entrada: Matriz $A \in \mathbb{R}^{n\times n}$.
Saída: Autovetores (matriz ortogonal U) e autovalores de A (diagonal da matriz A).
início
  $U \leftarrow 1$;
  para $k = 0, \ldots,$ até convergência **faça**
    $A = QR$ (decomposição QR);
    $A \leftarrow RQ$;
    $U \leftarrow UQ$;
  fim
fim

mamente simples, mas o mecanismo por trás do seu bom funcionamento é bastante sutil. Notem que o passo iterativo básico é

$$\begin{aligned} A_k &= Q_k R_k \\ A_{k+1} &= R_k Q_k = Q_k^t A_k Q^k. \end{aligned} \tag{34}$$

Por construção, o algoritmo, a cada passo, gera uma matriz A_{k+1} similar à anterior A_k e, portanto, que têm os mesmos autovalores. É fácil ver que começando-se com $A_0 = A$, teremos

$$A_{k+1} = \overbrace{Q_k^t \cdots Q_2^t Q_1^t}^{P_k^t} A \underbrace{Q_1 Q_2 \cdots Q_k}_{P_k}, \tag{35}$$

sendo P_k uma matriz ortogonal. O ponto central do algoritmos é que sempre temos

$$\lim_{k\to\infty} P_k = U \tag{36}$$

e, portanto, $\lim_{k\to\infty} A_k = D$, e o problema da diagonalização está resolvido. A questão, obviamente, está em provar (36). A literatura sobre esse assunto é vasta, há diversos detalhes sutis que não são importantes para o nosso caso de matrizes simétricas. Vamos mostrar que o algoritmo 1 funciona porque, basicamente, ele é equivalente a um algoritmo muito mais intuitivo, o da iteração de subespaços, que será exposto a seguir.

Considerem a sequência $\{X, AX, A^2X, \dots\}$ de vetores, com $X \in \mathbb{R}^n$ arbitrário. O que podemos afirmar sobre essa sequência para o caso de A simétrico? Bem, nesse caso, os autovetores u_k de A formam uma base ortonormal de $\mathbb{R}^n$, e teremos $X = \sum_{k=1}^{n} \alpha_k u_k$. Vamos supor os autovalores associados ordenados como $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$. O caso de eventuais autovalores repetidos não será um problema. Teremos

$$A^m X = \sum_{k=1}^{n} \alpha_k \lambda_k^m u_k = \lambda_1^m \left(\alpha_1 u_1 + \sum_{k=2}^{n} \alpha_k \left(\frac{\lambda_k}{\lambda_1} \right)^m u_k \right), \tag{37}$$

de onde conclui-se que

$$\lim_{m \to \infty} A^m X = \alpha_1 \lambda_1^m u_1, \tag{38}$$

quer dizer, a iteração da multiplicação por A "seleciona" a direção do autovetor de maior autovalor em módulo de A. Na verdade, seleciona a direção do maior autovetor de A no qual X tem componente, mas isto não muda a nossa análise. A divergência em (38) para $m \to \infty$ não é um problema, pois estamos interessados apenas nas direções. Seria possível propor uma sequencia que selecionasse, por exemplo, os dois autovetores u_1 e u_2 associados aos dois maiores autovalores em módulo? A resposta é sim, e basicamente devemos gerar duas sequências começando-se com dois vetores X e Y linearmente independentes. Porém, se simplesmente repetirmos as iterações, terminaremos com as duas sequências convergindo na mesma direção, a de u_1. Como garantir, por exemplo, que a sequencia de X convirja pra u_1 e a de Y convirja pra u_2? A resposta é: garantindo que as duas sequências sejam ortogonais entre si, isto é, $(A^k Y)^t (A^k X) = 0$ para todo k. Com isto, estaremos garantindo que a sequência $A^k X$ converge para a direção de u_1, enquanto a $A^k Y$ converge para a direção de máximo autovalor em módulo do subespaço ortogonal a u_1, quer dizer, estamos selecionando u_2 automaticamente. Se além de exigir ortogonalidade também impusermos que todos os vetores da sequência sejam unitários, evitaremos o problema da divergência em (38) para $m \to \infty$. Esta operação, na prática, corresponde a aplicar um procedimento de Gram-Schmidt a cada iteração.

Bem, temos, em princípio, uma ideia que pode ser explorada para propor sequências que convirjam para **todos** os autovetores de A. Basta construirmos n sequências ortonormais como as anteriores. Vamos usar como vetores iniciais as colunas da matriz identidade, o que já nos garante que os primeiros elementos das sequencias são ortogonais. Os próximos elementos na

sequencia são as próprias colunas de A (pois os próximos elementos são as colunas do produto de A com a identidade). Porém, estamos interessados apenas nos vetores ortogonais, então podemos fazer $A = QR$ e utilizar, no passo seguinte, a matriz AQ, que por sua vez seria novamente decomposta em um produto QR e assim sucessivamente. Estes passos correspondem ao algoritmo 2, chamado de iteração de subespaços, cuja convergência é muito

Algoritmo 2: Algoritmo da iteração de subespaços

Entrada: Matriz $A \in \mathbb{R}^{n \times n}$.
Saída: Autovetores (matriz ortogonal U) e autovalores de A (diagonal da matriz D).
início
  $U \leftarrow 1$;
  para $k = 0, \ldots,$ até convergência **faça**
    $AU = QR$ (decomposição QR);
    $U \leftarrow Q$;
  fim
  $D \leftarrow U^t AU$;
fim

mais intuitiva que a do algoritmo 1. A iteração básica aqui é

$$AU_k = U_{k+1} R_{k+1}. \tag{39}$$

Porém, aqui esperamos que[3] $\lim_{k \to \infty} U_k = U$, quer dizer, U_k converge para a a matriz ortogonal dos autovetores de A.

Notem que, combinando-se (34) e (35), temos que o passo do algoritmo QR pode ser escrito como

$$Q_k^t \cdots Q_2^t Q_1^t A Q_1 Q_2 \cdots Q_k = Q_{k+1} R_{k+1}, \tag{40}$$

de onde temos

$$A \underbrace{Q_1 Q_2 \cdots Q_k}_{U_k} = \underbrace{Q_1 Q_2 \cdots Q_k Q_{k+1}}_{U_{k+1}} R_{k+1}, \tag{41}$$

comprovando que os algoritmos 1 e 2 são, na prática, idênticos, estabelecendo a convergência de 1 com base na convergência esperada de 2.

[3]Não provamos rigorosamente, mas a prova para o caso simétrico segue as linhas do que foi exposto até aqui.

Em princípio, temos tudo o que precisamos para implementar um algoritmo bastante decente de diagonalização e, com ele, podemos implementar um de SVD. Há, porém, um ultimo passo que vale a pena expor. Há ganhos consideráveis, tanto de desempenho como de precisão, se aplicarmos o algoritmo QR a uma redução tridiagonal de A. Isto significa que não aplicaremos o algoritmo QR diretamente na matriz A, mas sim numa matriz semelhante Q^tAQ que seja tridiagonal, *i.e.*, tem todas as entradas nulas exceto a diagonal principal, uma diagonal acima e uma abaixo. Não é difícil determinar que matriz ortogonal Q reduz uma matriz simétrica A a uma forma tridiagonal, fica de exercício. A dica é usar as transformações de Householder.

A simplificação vem do fato que o caráter tridiagonal da matriz A será mantido ao longo das iterações. É interessante tentar visualizar o que acontece. A decomposição QR de uma matriz tridiagonal será do tipo

$$\overbrace{\begin{pmatrix} \bullet & \bullet & & & \\ \bullet & \bullet & \bullet & & \\ & \bullet & \bullet & \bullet & \\ & & \bullet & \bullet & \bullet \\ & & & \bullet & \bullet \end{pmatrix}}^{A} = \overbrace{\begin{pmatrix} \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ & \bullet & \bullet & \bullet & \bullet \\ & & \bullet & \bullet & \bullet \\ & & & \bullet & \bullet \end{pmatrix}}^{Q} \overbrace{\begin{pmatrix} \bullet & \bullet & \bullet & & \\ & \bullet & \bullet & \bullet & \\ & & \bullet & \bullet & \bullet \\ & & & \bullet & \bullet \\ & & & & \bullet \end{pmatrix}}^{R}. \tag{42}$$

As entradas de Q abaixo da segunda diagonal são nulas porque a coluna k de Q é a coluna k de A ortogonalizada (Gram-Schmidt) com relação às anteriores e, portanto, da natureza triangular de A teremos $(Q)_{ij} = 0$ para $i > j + 1$. Para determinarmos a forma de R, notem que $R = Q^*A$

$$\overbrace{\begin{pmatrix} \bullet & \bullet & \bullet & & \\ & \bullet & \bullet & \bullet & \\ & & \bullet & \bullet & \bullet \\ & & & \bullet & \bullet \\ & & & & \bullet \end{pmatrix}}^{R} = \overbrace{\begin{pmatrix} \bullet & \bullet & & & \\ \bullet & \bullet & \bullet & & \\ \bullet & \bullet & \bullet & \bullet & \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \end{pmatrix}}^{Q*} \overbrace{\begin{pmatrix} \bullet & \bullet & & & \\ \bullet & \bullet & \bullet & & \\ & \bullet & \bullet & \bullet & \\ & & \bullet & \bullet & \bullet \\ & & & \bullet & \bullet \end{pmatrix}}^{A}, \tag{43}$$

de onde é fácil ver que $(R)_{ij} = 0$ para $j > i + 2$. O passo do algoritmo QR é o

cálculo do produto RQ. Neste caso, teremos

$$\tilde{A} = \overbrace{\begin{pmatrix} \bullet & \bullet & \bullet & & \\ & \bullet & \bullet & \bullet & \\ & & \bullet & \bullet & \bullet \\ & & & \bullet & \bullet \\ & & & & \bullet \end{pmatrix}}^{R} \overbrace{\begin{pmatrix} \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ & \bullet & \bullet & \bullet & \bullet \\ & & \bullet & \bullet & \bullet \\ & & & \bullet & \bullet \end{pmatrix}}^{Q}. \tag{44}$$

É fácil ver que $(\tilde{A})_{ij} = 0$ para $i > j + 1$. Porém, como $\tilde{A}$ é simétrica, temos que será também tridiagonal.

Já chegaram relatos de que, provavelmente por erros numéricos, a matriz RQ não é exatamente tridiagonal. Isto pode (deve) ser evitado calculando-se **apenas** os elementos que sabemos ser não nulos, impondo-se que todos os outros sejam zero. De fato, em aplicações nas quais a questão de memória é critica, apenas as entradas sabidamente não zero são armazenadas.

Referências

[1] G.W. Stewart, *On the Early History of the Singular Value Decomposition*, SIAM Review **35**, 551 (1993).

[2] C.F. Van Loan, *The ubiquitous Kronecker product*, J. Comput. Appl. Math. **123**, 85 (2000).

[3] R. Horn, C. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, (1991)

[4] W. Steeb e Y. Hardy, *Entangled Quantum States and the Kronecker Product*, Zeitschr. Naturforschung A **57**, 689 (2002).

[5] W.T. Vetterling, B.P. Flannery, W.H. Press, S. Teukolsky, *Numerical Recipes: The Art of Scientific Computing*, (2007).