CISC 271 Class 17

Matrix Approximation

Text Correspondence, Strang 2019: pp. 71–75 [14]

Main Concepts:

- *Matrix norm: extension of vector norm*
- *SVD as a matrix series*
- *Matrix approximation as a truncated series*

Sample Problem, Machine Inference: What is the "best" approximation of a matrix?

One of the many uses of the SVD is that its formulation produces a *series* for any given matrix [3]. To understand how we can use such a series in linear data analysis, we will first explore the norm of a matrix of a matrix. ℓ^2

17.1 Matrix Norms

A *norm* of a vector or matrix is a real number that "measures" the object. We write the vector norm as $\|\vec{a}\|$ and the matrix norm as $\|A\|$; a common abbreviation for the norm of an object is $\|\cdot\|$. A norm must satisfy four axioms, which we will write for a matrix. The first four axioms are the same as the axioms for a vector norm and some authors add a fifth axiom for matrix norms: the object being measured has changed.

For any $A \in \mathbb{R}^{m\times n}$, any $A \in \mathbb{R}^{m\times n}$, and any $\alpha \in \mathbb{R}$, the axioms of a norm are:

- $\|A\| \geq 0$
- $\|A\| = 0$ if and only if $A = 0$
- $\|\alpha A\| = |\alpha|\ \|A\|$
- $\|A + C\| \leq \|A\| + \|C\|$
- $\|AC\| \leq \|A\|\ \|C\|$ (not universally required)

Sometimes, we want a vector norm and a matrix norm to "work" together.

Definition: compatible norms

For any vector norm $\|\cdot\|$,any matrix norm $\|\cdot\|$, any vector $\vec{w} \in \mathbb{R}^n$, and any matrix $A \in \mathbb{R}^{m\times n}$, the norms are *compatible* is defined as

$$\|A\vec{w}\| \leq \|A\|\ \|\vec{w}\| \tag{17.1}$$

The vector norm that we use in this course is the Euclidean norm. It is also written as the ℓ_2 norm, pronounced *ell–2*, which is the abbreviation that we will prefer. For a vector $\vec{w} \in \mathbb{R}^n$, this norm is defined as

$$\|\vec{w}\|_2 \stackrel{\text{def}}{=} \sqrt{\sum_{j=1}^{n} (w_j)^2} \tag{17.2}$$

The ℓ_2 matrix norm is an *induced* norm, which means that it is based on the vector norm. For a matrix $A \in \mathbb{R}^{m \times n}$, the ℓ_2 norm is

$$\|A\|_2 \stackrel{\text{def}}{=} \max_{\vec{w} \neq \vec{0}} \frac{\|A\vec{w}\|}{\|\vec{w}\|} \tag{17.3}$$

The *Frobenius* norm is another matrix norm that is closely related to the Euclidean vectorm norm. It is defined as

$$\|A\|_F \stackrel{\text{def}}{=} \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} (a_{ij})^2} \tag{17.4}$$

17.1.1 Eigenvalues and Singular Values

The above norms are closely related to matrix decompositions that we explored earlier in the course. These relations are theorems in linear algebra that can be found in many textbooks and other sources.

Abbreviate the largest eigenvalue of a matrix as $\lambda_{MAX}(\cdot)$ and recall the SVD of a matrix, in which the singular values are ordered as $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_r > 0$. The above matrix norms have the properties

$$\begin{aligned} \|A\|_2 &= \sqrt{\lambda_{MAX}(A^T A)} \\ &= \sigma_1 \\ \|A\|_F &= \sqrt{\sum_{j=1}^{r} \sigma_j^2} \end{aligned} \tag{17.5}$$

17.2 A Matrix As A Series

Let us recall the SVD of a matrix $A \in \mathbb{R}^{m \times n}$ that is rank r. Because we can neglect each row and each column that has an index greater that r, we can write this matrix as

$$\begin{aligned} A &= U\Sigma V^T \\ \text{where} \quad U &= \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_r \end{bmatrix} \\ \Sigma &= \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix} \\ V^T &= \begin{bmatrix} \vec{v}_1^T \\ \vec{v}_2^T \\ \vdots \\ \vec{v}_r^T \end{bmatrix} \end{aligned} \tag{17.6}$$

We can expand the right product of Equation 17.6 as

$$\begin{aligned} \Sigma V^T &= \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix} \begin{bmatrix} \vec{v}_1^T \\ \vec{v}_2^T \\ \vdots \\ \vec{v}_r^T \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1 \vec{v}_1^T \\ \sigma_2 \vec{v}_2^T \\ \vdots \\ \sigma_r \vec{v}_r^T \end{bmatrix} \end{aligned} \tag{17.7}$$

Using Equation 17.7, we can write the SVD of Equation 17.6 as

$$\begin{aligned} A &= U\Sigma V^T \\ &= \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_r \end{bmatrix} \begin{bmatrix} \sigma_1 \vec{v}_1^T \\ \sigma_2 \vec{v}_2^T \\ \vdots \\ \sigma_r \vec{v}_r^T \end{bmatrix} \\ &= \vec{u}_1 \sigma_1 \vec{v}_1^T + \vec{u}_2 \sigma_2 \vec{v}_2^T + \cdots + \vec{u}_r \sigma_r \vec{v}_r^T \\ &= \sigma_1 [\vec{u}_1 \vec{v}_1^T] + \sigma_2 [\vec{u}_2 \vec{v}_2^T] + \cdots + \sigma_r [\vec{u}_r \vec{v}_r^T] \end{aligned} \tag{17.8}$$

17.3 Matrix Approximation: Eckart-Young Theorem

Equation 17.8 is a remarkable series. It states that any rank–r matrix A can be represented as the sum of r rank–1 matrices. Each rank–1 matrix in the series is the product of the corresponding left singular vector and right singular vector of A.

The SVD was observed and discovered, at least partially, since the late 19^{th} century. It was formulated and shown to exist in 1936, by Eckart and Young [3]; their approximation is known in linear algebra as the Eckart-Young Theorem.

The approximation theorem has been stated and proved for many matrix norms. We can use either the ℓ_2 norm or the Frobenius norm. For a matrix A, we will define the i^{th} rank–1 matrix in Equation 17.8 as

$$C_i \stackrel{\text{def}}{=} \sigma_i[\vec{u}_i \vec{v}_i^T] \tag{17.9}$$

Formally, the Eckart-Young theorem states that the optimal rank–k approximation to the matrix A is the first k terms of the series in Equation 17.8. For $A \in \mathbb{R}^{m \times n}$, the optimal approximation is

$$\begin{aligned} C &= \underset{W \in \mathbb{R}^{m \times n}}{\operatorname{argmin}} \|A - W\|_2 \\ &= \underset{W \in \mathbb{R}^{m \times n}}{\operatorname{argmin}} \|A - W\|_F \\ &= C_1 + C_2 + \cdots + C_k \end{aligned} \tag{17.10}$$

Using Equation 17.10 and Equation 17.6, we can find the SVD of the optimal rank–k approximation to A as

$$C = U_k \Sigma_k V_k^T \tag{17.11}$$

$$\text{where} \quad U_k = \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_k \end{bmatrix}$$

$$\Sigma_k = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_k \end{bmatrix}$$

$$V_k^T = \begin{bmatrix} \vec{v}_1^T \\ \vec{v}_2^T \\ \vdots \\ \vec{v}_k^T \end{bmatrix}$$

We use Equation 17.11 often when we perform linear data analysis. One way that we can think of this approximation is:

The optimal rank–k approximation to the column space of A is U_k

That is, when we want to use a smaller set of basis vectors to span the column space of a matrix A, the "best" choice is the first k left singular vectors of the matrix.

17.4 Approximations and The Scree Plot

How can we select the rank$k$ of an approximation? In practice, there is no universally accepted method. There may be guidelines within a domain of application but, in general, we usually examine the data and exercise human judgement.

A common method for determining a matrix approximation is the use of a *scree plot* [2]. Catell devised the name by a visual analogy to "... the straight line of rubble and boulders which forms at the pitch of sliding stability at the foot of a mountain".

In terms of the SVD, a scree plot has a horizontal axis, or ordinate, the is the integers i of Equation 17.9, ranging from 1 to either r or a suitable number less than r. The vertical axis, or abscissa, is a normalized version of the singular value σ_i. The normalization is typically the sum of the singular values, which is the explained variance θ, or the square root of the sum of the squares of the singular values, which is the total variance T:

$$\theta = \sum_{i=1}^{r} \sigma_i \tag{17.12}$$

$$T = \sum_{i=1}^{r} \sigma_i^2 \tag{17.13}$$

The choice of θ from Equation 17.12, or either T or $\sqrt{T}$ from Equation 17.13, depends on the application.

A simple example can be developed by generating a random matrix of integers. For one such matrix, which is $A \in \mathbb{R}^{20\times 20}$ with $a_{ij} \leq 200$, the scree plot is shown in Figure 17.1. In this example the normalization was T from Equation 17.13. The scree plot suggests that the 20×20 matrix has a column space that can be approximated with 2 basis vectors, because the index 2 is where the scree plot begins to level.

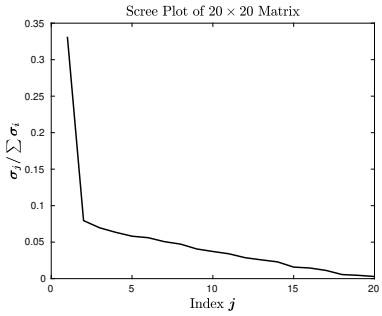

Figure 17.1: Scree plot of a random 20×20 integer-valued matrix. The index $k = 2$ would be chosen as the rank of approximation of this matrix.

The scree plot is one way that we can use the Eckart-Young theorem to select the number of basis vectors when we want to approximate the column space of data. These basis vectors are part of *principal component analysis*, which we will explore after we better understand vector spaces and projections.