9

Singular Value Decomposition

9.1 Introduction

In many modern applications involving large data sets, statisticians are confronted with a large \( m \times n \) matrix \( X = (x_{ij}) \) that encodes \( n \) features on each of \( m \) objects. For instance, in gene microarray studies \( x_{ij} \) represents the expression level of the \( i \)th gene under the \( j \)th experimental condition [13]. In information retrieval, \( x_{ij} \) represents the frequency of the \( j \)th word or term in the \( i \)th document [2]. The singular value decomposition (svd) captures the structure of such matrices. In many applications there are alternatives to the svd, but these are seldom as informative or as numerically accurate.

Most readers are well acquainted with the spectral theorem for symmetric matrices. This classical result states that an \( m \times m \) symmetric matrix \( A \) can be written as \( A = U \Sigma U^t \) for an orthogonal matrix \( U \) and a diagonal matrix \( \Sigma \) with diagonal entries \( \sigma_i \). If \( U \) has columns \( u_1, \ldots, u_m \), then the matrix product \( A = U \Sigma U^t \) unfolds into the sum of outer products

\[
A = \sum_{j=1}^{m} \sigma_j u_j u_j^t .
\]

When \( \sigma_j = 0 \) for \( j > k \), \( A \) has rank \( k \) and only the first \( k \) terms of the sum are relevant. The svd seeks to generalize the spectral theorem to nonsymmetric matrices. In this case there are two orthonormal sets of vectors \( u_1, \ldots, u_k \) and \( v_1, \ldots, v_k \) instead of one, and we write

\[
A = \sum_{j=1}^{k} \sigma_j u_j v_j^t = U \Sigma V^t \tag{9.1}
\]

for matrices \( U \) and \( V \) with orthonormal columns \( u_1, \ldots, u_k \) and \( v_1, \ldots, v_k \), respectively. Remarkably, it is always possible to find such a representation with nonnegative \( \sigma_j \).

For some purposes, it is better to fill out the matrices \( U \) and \( V \) to full orthogonal matrices. If \( A \) is \( m \times n \), then \( U \) is viewed as \( m \times m \), \( \Sigma \) as \( m \times n \),
and $V$ as $n \times n$. The svd then becomes

$$A = (u_1 \ldots u_k u_{k+1} \ldots u_m) \begin{pmatrix} \sigma_1 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_k & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & 0 & \ldots & 0 \end{pmatrix} \begin{pmatrix} v_1^t \\ \vdots \\ v_k^t \\ v_{k+1}^t \\ \vdots \\ v_n^t \end{pmatrix},$$

assuming $k < \min\{m, n\}$. The scalars $\sigma_1, \ldots, \sigma_k$ are said to be singular values and conventionally are listed in decreasing order. The vectors $u_1, \ldots, u_k$ are known as left singular vectors and the vectors $v_1, \ldots, v_k$ as right singular vectors. We will refer to any svd with $U$ and $V$ orthogonal matrices as full.

The numerical issues in constructing the svd are so complicated that outsiders get the impression that only experts should write code. Fortunately, the experts are kind enough to contribute open source code to cooperative efforts such as LAPACK [1]. It would be a mistake not to take advantage of these efforts. However, perfectly usable software can be written by novices for small-scale projects. Our discussion of Jacobi’s method for constructing the svd targets such readers. Because it is accurate and receptive to parallelization, Jacobi’s method remains in contention with other algorithms. In addition to these advantages, it is easy to explain.

There is a huge literature on the svd. The two books [9, 10] by Horn and Johnson provide an excellent overview of the mathematical theory. For all things numerical, the treatise of Golub and Van Loan [7] is the definitive source. Beginners will appreciate the more leisurely pace of the books [6, 15, 17]. At the same level as this text, the books [4, 16, 18] are also highly recommended.

### 9.2 Basic Properties of the SVD

Let us start with the crucial issue of existence following Karro and Li [12].

**Proposition 9.2.1** Every $m \times n$ matrix $A$ has a singular value decomposition of the form (9.1) with positive diagonal entries for $\Sigma$.

**Proof:** It suffices to prove that $A$ can be represented as $U\Sigma V^t$ for full orthogonal matrices $U$ and $V$. We proceed by induction on $\min\{m, n\}$. The cases $m = 1$ and $n = 1$ are trivial. In the first case, we write $A = \sigma_1 v^t$, and in the second case $A = \sigma u_1$. Thus, assume that $m > 1$ and $n > 1$. According to Proposition 11.2.1, the continuous function $(x, y) \mapsto x^tAy$ attains its maximum $\sigma_1$ on the compact set

$$S = \{(x, y) : x^tx = y^ty = 1\}$$