Discussion 12 **Alex Francis**
CS 189 Email: afrancis@berkeley.edu
04/17/2017 OH: Wednesday 10:00am - 12:00pm (283E Soda)

1 Lecture on the Singular Value Decomposition

Matrix factorizations are useful for illuminating hidden structure in matrix equations. The Spectral Decomposition is one such factorization, and is relevant to some matrix $A \in \mathbb{S}^{n \times n}$, i.e. to symmetric $n \times n$ matrices. Recall that the Spectral Decomposition has the form,

$$A = U \Lambda U^\top$$

The Spectral Theorem, stated and proved (out of scope) below, tells us that any symmetric matrix has a decomposition of that form.

THEOREM 1.1 **(The Spectral Theorem)** Let $A \in \mathbb{R}^{n \times n}$ be symmetric, let $\lambda_i \in \mathbb{R}, i = 1, \ldots, n$, be the eigenvalues of A (counting multiplicities). Then, there exists a set of orthonormal vectors $u_i \in \mathbb{R}^n, u = 1, \ldots, n$ such that $Au_i = \lambda_i u_i$. Equivalently, there exist an orthogonal matrix $U = \begin{bmatrix} u_1 \ldots u_n \end{bmatrix}$ (i.e., $UU^\top = U^\top U = I_n$) such that,

$$A = U \Lambda U^\top = \sum_{i=1}^{n} \lambda_i u_i u_i^\top, \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$$

Proof. Here's a simple and non-rigorous argument (establishing the existence of n eigenvalues corresponding to a set of n eigenvectors that form basis of $\mathbb{R}^n$ is non-trivial). The remainder of the proof follows the sketch of Dasgupta (one of the authors of the CS 170 textbook) in `https://cseweb.ucsd.edu/~dasgupta/291-unsup/lec7.pdf`. Note that the matrix U represents a linear transformation, and is completely determined by its behavior on any set of n linearly independent vectors. It is therefore true that if $Au_i = Mu_i$ for any $M \in \mathbb{R}^{n \times n}$ and for all $i \in [n]$, that A and M are identical. Then, for any i,

$$U \Lambda U^\top u_i = U \Lambda e_i = U \lambda_i e_i = \lambda_i U e_i = \lambda_i u_i = A u_i$$

In the above, e_i is the standard ith basis vector. The proof is complete. □

We have seen throughout this course that the Spectral Decomposition is useful for establishing certain theoretical results in machine learning. For example, the Spectral Decomposition can be used to prove that the maximum and minimum of the Rayleigh quotient is, in fact, that maximum and minimum eigenvalue of a particular symmetric matrix. The Spectral Decomposition is also useful for defining the matrix square-root via the Cholesky decomposition.

The Singular Value Decomposition extends the concept of matrix factorizations to non-symmetric $m \times n$ matrices. The form of this factorization is familiar, and the so-called *SVD Theorem* is stated below.

THEOREM 1.2 **(The SVD Theorem)** Any matrix $A \in \mathbb{R}^{m \times n}$ can be factored as,

$$A = U \tilde{\Sigma} V^\top$$

Where $V \in \mathbb{R}^{n\times n}$ and $U \in \mathbb{R}^{m\times m}$ are orthogonal matrices, and $\tilde{\Sigma} \in \mathbb{R}^{m\times n}$ is a matrix having the first $r = \text{rank}A$ diagonal entries $(\sigma_1, \ldots, \sigma_r)$ positive and decreasing in magnitude, and all other entries zero:

$$\tilde{\Sigma} = \begin{bmatrix} \Sigma & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{bmatrix} = \sum_{i=1}^{r} \sigma_i u_i v_i^\top, \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r)$$

Where $\sigma_i^2 = \lambda_i(AA^\top) = \lambda_i(A^\top A)$ and u_i are the eigenvectors of $A^\top A$ and v_i are the eigenvectors are $AA^\top$. The proof is fairly technical, and out of scope. It relies on the Spectral Decomposition for Symmetric matrices. The Singular Value Decomposition is useful for proving certain properties of non-symmetric matrices. However, perhaps more importantly, it has an alternative interpretation that is quite useful for PCA.

Application 1. Low-Rank Matrix Approximation and Minimum Distance to Rank Deficiency

Suppose $A \in \mathbb{R}^{m\times n}$ is a given matrix with $\text{rank}A = r > 0$. We consider the problem of approximating A with a matrix of lower rank. In particular, we consider the following rank-constrained approximation problem,

$$\min_{A_k \in \mathbb{R}^{m\times n}} \|A - A_k\|_F^2$$
$$\text{subject to: } \text{rank}(A_k) = k$$

The optimal solution to this problem is just the Singular Value decomposition truncated to the first k terms. This tells us that the k-dimensional subspace that best represents the data (the "maximum-variance projection") is just the rank k version of the Singular Value Decomposition. Recall that this is *precisely* what we were interested in doing in the PCA algorithm - projecting the data onto the most representative lower-dimensional subspace! Therefore, we have the PCA algorithm... (1) Compute the SVD, and (2) Project the points onto lower-dimensional subspaces by using the SVD - see Shewchuk's notes.