Stat 991: Multivariate Analysis, Dimensionality Reduction, and Spectral Methods **Lecture: 1**

The Singular Value Decomposition

Instructor: Sham Kakade

1 Intro

The SVD is the single most important concept to understand in linear algebra. Intuitively, it precisely characterizes a way to view how any linear map behaves.

Roughly speaking, the SVD corresponds to a certain natural notion of "geometric" regression. In fact, with this interpretation, all of classical estimation issues (with noisy data) are relevant here.

1.1 Vanilla Regression and a "Best Fit Line"

Consider an input data matrix $X_{\text{in}} \in \mathbb{R}^{n \times d}$ and our target prediction vector X_{out}. In regression, we desire to predict the target with the inputs. In a least squares sense, the goal is to find w which minimizes:

$$\min_w \|X_{\text{out}} - X_{\text{in}} w\|^2$$

Here, the solution is given by:

$$w = (X_{\text{in}}^\top X_{\text{in}})^{-1} X_{\text{out}}$$

This is the least squares estimator.

Question 1. *With noisy data, how accurate is our regression?*

1. *(fixed design) when X_{out} is random and X_{in} is fixed?*
2. *(random design) when X_{out} and X_{in} are random?*

2 The Best Fit Line, Rotationally Invariant Regression, and Matrix Norms

2.1 The Best Fit Line

In "vanilla" regression, note that there was one preferred coordinate which we desired to predict, and we fit our data with a line. Instead, let us say we have no preferred direction (with which to measure our error), and, yet, we still desire to fit our data with a line. In particular, this can be viewed as a rotationally invariant, geometric generalization of regression — precisely, what is the best fit line to our data, measured with respect to the rationally invariant Euclidean norm.

Note that for any vector x, the best fit point on our line w is $\frac{w \cdot x}{\|w\|^2} w$. Without loss of generality, let us constrain w to be unit norm, i.e. $\|w\| = 1$.

Now let $X \in \mathbb{R}^{n \times d}$. Let us consider fitting the best a line to the rows $x_i \in \mathbb{R}^d$ of this matrix. Hence, the best fit line w is the solution to the problem:

$$\min_w \sum_i \|x_i - (w \cdot x_i) w\|^2$$

(where $w \in \mathbb{R}^d$). Equivalently, we can find w as a solution to the maximization problem in the following lemma:

Lemma 2.1. *We have that:*

$$\sum_i \|x_i - (w \cdot x_i)w\|^2 = \sum_i \|x\|^2 - \sum_i (w \cdot x_i)^2 = \|X\|_F - \|Xw\|^2$$

where $\|X\|_F$ is the Frobenius norm *(e.g. the sum of the squares of the entries). Hence, the best fit line is given by:*

$$\arg \max_{w:\ \|w\|=1} \|Xw\|^2 ..$$

Now one key step in understanding the SVD (presented later) is understanding the answer to the following question:

Question 2. *Let v be the best fit line to the rows of X. What is the best fit line to the columns of X, as a function of v and X?*

To answer this, let us first examine some norms.

2.2 The Spectral Norm and a little duality

The *spectral norm* $\|X\|$ of a matrix X is defined as:

$$\|X\| = \max_{a:\ \|a\|=1} \|Xa\|$$

Note it is *rotationally invariant*.

Perhaps some intuition for this norm can be obtained by viewing the Euclidean norm as a certain maximization problem: we can write the Euclidean norm of a vector a as:

$$\|a\| = \sqrt{a \cdot a} = \max_{b:\|b\|=1} b \cdot a$$

which follows from Cauchy-Schwartz.

To understand the previous question (of the best fit line to the columns of X), observe that:

$$\|X\| = \max_{b:\|b\|=1} b \cdot a = \max_{a,b:\|a\|=\|b\|=1} b^\top X a \tag{1}$$

where $b \in \mathbb{R}^n$ and $a \in \mathbb{R}^d$.

2.3 The Best Line of the Columns

Lemma 2.2. *If v is the best fit line to the columns of X, then Xv is the best fit line to the rows of X.*

Proof. Observe that if v is the argmax:

$$v = \arg \max_{a:\ \|a\|=1} \|Xa\| = \arg \max_{a:\ \|a\|=1} \left(\max_{b:\|b\|=1} b^\top X a \right)$$

Note that the b which achieves the max must be $\frac{Xv}{\|Xv\|}$. This is because $\frac{a}{\|a\|}$ is the (unit length) b which maximizes $b \cdot a$.

Hence,the argmax over (u, v) in Equation 1 is achieved by:

$$(\frac{Xv}{\|Xv\|}, v) = \arg \max_{a,b:\|a\|=\|b\|=1} b^\top X a$$

Note this implies that:

$$\frac{Xv}{\|Xv\|} = \arg \max_{b:\|b\|=1} \left(\max_{a:\|a\|=1} u^\top X v \right)$$

Equivalently,

$$\frac{Xv}{\|Xv\|} = \arg \max_{b:\|b\|=1} \left(\max_{a:\|a\|=1} b^\top X^\top a \right) = \arg \max_{b:\|b\|=1} \|X^\top b\|$$

Hence, $\frac{Xv}{\|Xv\|}$ specifies the best fit line to the columns of X. So Xv is also a best fit line (though not necessarily of unit length). □

3 The Best Fitting Subspace and the SVD

Now we let us X be a general matrix. The maximal *singular value* is $\max_{\|w\|=1} \|Av\|_2$ and the argmax is the corresponding singular vector. We let A_i be a row of A.

Lemma 3.1. *For an arbitrary matrix $A \in \mathbb{R}^{n \times d}$,*

$$\arg\max_{\|w\|=1} \|Aw\|^2 = \arg\min_{\|w\|=1} \|A - (Aw)w^\top\|_F^2 = \arg\min_{\|w\|=1} \sum_i \|A_i - (A_i \cdot w)w\|^2$$

where $\|\cdot\|_F^2$ is the Frobenius norm (the Frobenius norm of a matrix M is $\|\cdot\|_F^2 = \sum_{i,j} M_{i,j}^2$).

Proof. The proof essentially follows from the Pythagoras theorem. □

Theorem 3.2. *(SVD) Define the k dimensional subspace V_k as the span of the following k vectors:*

$$v_1 = \arg\max_{\|v\|=1} \|Av\|^2 \tag{2}$$

$$v_2 = \arg\max_{\|v\|=1, v\cdot v_1=0} \|Av\|^2 \tag{3}$$

$$\vdots \tag{4}$$

$$v_k = \arg\max_{\|v\|=1, \forall i \le k,\, v\cdot v_i=0 \forall i \le k} \|Av\|^2 \tag{5}$$

Then V_k is optimal in the sense that:

$$V_k = \arg\min_{dim(V)=k} \sum_i distance(A_i, V_k)^2$$

Furthermore,

$$\sigma_1 = \|Av_1\| \ge \sigma_2 = \|Av_2\| \ge \ldots \sigma_{\min\{n,d\}} = \|Av_{\min\{n,d\}}\|$$

Let $\sigma_i u_i = Av_i$, so u_i is unit length. Then the set $\{u_i\}$ is orthonormal (so is $\{v_i\}$ by construction) and the SVD decomposition of A is:

$$A = \sum_i \sigma_i u_i v_i^\top = U diag(\sigma_1, \ldots \sigma_{\min\{n,d\}}) V^\top$$

where U and V are orthogonal matrices with rows $\{u_i\}$ and $\{v_i\}$, respectively.

Proof. The interesting part of the proof is that $\{u_i\}$ is orthonormal — the rest of the proof essentially follows by construction. □

As a corollary, we have that:

Corollary 3.3. *Among all rank k matrices D, $A_k = \Sigma_{i=1}^k \sigma_i u_i v_i^\top$ is the one which minimizes $\|A - D\|_F$. Furthermore,*

$$\|X - D\|_F^2 = \|X\|_F^2 - \sum_{i=1}^k \sigma_i^2 = \sum_{i=k+1}^{\min\{n,d\}} \sigma_i^2$$

3.1 Proofs

The argument is essentially an inductive one based on the previous argument.

3.2 Three Interpretations

The relevance of the SVD is that it holds for all matrices (e.g. it's a characterization of all linear maps).

1. The SVD shows that any linear map consists of a rotation followed by an (axis aligned) scaling followed by another rotation.
2. The best fit k-dimensional subspace to the rows is V_k. Furthermore, the best fit $k+1$-dimensional subspace contains the best fit k dimensional subspace (even if there are equal singular values, we can always choose subspaces such that this holds).
3. The best fit k-dimensional subspace is specified by the span of $Xv_1, \ldots Xv_k$.

4 References

Material used was Wikipedia and Santosh Vempala's lecture notes.