SVD, PCA and Metric Scaling

These notes provide a more formal treatment than the lectures, and prove all the linear mathematics used.

1 Singular Value Decomposition

Suppose we have a $n \times p$ matrix X. Where necessary we will assume that $n \geqslant p$ to ease the notation, but this is unnecessary. The Frobenius norm of X, $\|X\|$, is the square root of the sum of squares of the elements (and so the squared norm is the sum of the squared lengths of the rows or columns).

Proposition 1 *A $n \times p$ matrix X has a singular value decomposition of the form*

$$X = U \Lambda V^T$$

where Λ is a diagonal matrix with decreasing non-negative entries, U is a $n \times p$ matrix with orthonormal columns, and V is a $p \times p$ orthogonal matrix.

PROOF: Let λ_1 be the maximal length of Xx for a unit-length vector x, and let x and y be unit-length vectors such that $Xx = \lambda_1 y$. Extend y and x to orthogonal bases of $\mathbb{R}^n$ and $\mathbb{R}^p$ forming the columns of matrices U and V respectively. Then if

$$U = [y\ U_1], \qquad V = [x\ V_1]$$

we have, for $w^T = y^T X V_1$,

$$Y = U^T X V = \begin{bmatrix} y^T \\ U_1^T \end{bmatrix} X [x\ V_1] = \begin{bmatrix} \lambda_1 & w^T \\ 0 & X_1 \end{bmatrix}.$$

Since

$$\left\| Y \begin{pmatrix} \lambda_1 \\ w \end{pmatrix} \right\|^2 = \left\| \begin{pmatrix} \lambda_1^2 + \|w\|^2 \\ X_1 w \end{pmatrix} \right\|^2 \geqslant [\lambda_1^2 + \|w\|^2]^2$$

it follows that

$$\lambda_1^2 = \|X\|_2^2 = \|Y\|_2^2 \geqslant [\lambda_1^2 + \|w\|^2]$$

and so we must have $w = 0$. Now apply the argument inductively to X_1. □

Proposition 2 *Consider a $n \times p$ matrix X with singular value decomposition $X = U \Lambda V^T$. The best approximation in Frobenius norm to X by a matrix of rank $k \leqslant \min(n, p)$ is given by*

$$\widetilde{X} = U \mathrm{diag}(\lambda_1, \dots \lambda_k, \dots, 0) V^T.$$

This is also the best approximation by a projection[1] onto a subspace of dimension at most k, the projection onto the space spanned by the first k columns of U, and maximizes the Frobenius norm of a projection of X onto a subspace of dimension at most k.

PROOF: We have

$$\|X - \widetilde{X}\|^2 = \|\Lambda - \Lambda_k\|^2 = \sum_{k+1}^{\min(n,p)} \lambda_i^2.$$

$\widetilde{X}$ corresponds to a projection onto the space spanned by the first k columns of U, say U_k, since that projection gives

$$U_k(U_k^T U_k)^{-1} U_k^T X = U_k U_k^T U \Lambda V^T = U_k \Lambda_k V^T = U \Lambda_k V^T.$$

Consider any approximation Y of rank at most k. This can be written as $Y = AB$ where A is $n \times k$ and B is $k \times p$ (for example, via the SVD of Y). Now consider the best approximation of the form AC for any $k \times p$ matrix C. Since the squared Frobenius norm is the sum of the squared lengths of the columns, this is solved by regressing each column of X in turn on A; the optimal choice is $\widehat{C} = (A^T A)^{-1} A^T X$ and

$$\|X - Y\|^2 \geqslant \|X - A\widehat{C}\|^2 = \|[I - P_A]X\|^2 = \|X\|^2 - \|P_A X\|^2$$

where $P_A = A(A^T A)^{-1} A^T$ is the projection matrix onto span(A). Now we choose P_A to maximize $\|P_A X\|^2$:

$$\|P_A X\|^2 = \|P_A U \Lambda\|^2 = \sum_{1}^{\min(n,p)} \lambda_j^2 \|P_A u_j\|^2 = \sum_{1}^{\min(n,p)} \lambda_j^2 p_j^2$$

and $|p_j| \leqslant 1$ (it is the projection of a unit-length vector), $\sum p_j^2 = \|P_A U\|^2 = \|P_A\|^2 = k$. It is then obvious that the maximum is attained if and only if the first k p_j's are one, the rest zero, so

$$\|X - Y\|^2 \geqslant \|X\|^2 - \|P_A X\|^2 \geqslant \|X\|^2 - \sum_{1}^{k} \lambda_i^2 = \sum_{k+1}^{\min(n,p)} \lambda_i^2 = \|X - \widetilde{X}\|^2.$$

Any projection of X onto a subspace of k dimensions has rank at most k. □

It may help to note that projecting onto a subspace of dimension $k \leqslant p$ is equivalent to choosing an orthonormal $p \times k$ matrix A of linear combinations of the variables. Let $\mathbf{x}$ be a *row* vector denoting an observation, and let A be an arbitrary $p \times k$ matrix of full rank. We want to project onto the subspace spanned by the new variables, $\{Ay \mid y \in \mathbb{R}^k\}$. This is a regression problem, and the closest point to $\mathbf{x}^T$ is $(A^T A)^{-1} A^T \mathbf{x}^T$; so the projection corresponds to the matrix $A(A^T A)^{-1}$. This is an orthonormal matrix, and equal to A if it is itself orthonormal. Thus searching over all projections is equivalent to considering XA for all orthonormal A.

If X is a matrix whose rows are observations, proposition 2 gives:

Proposition 3 *Consider n p-variate observations forming a matrix X. Then the projection of proposition 2:*
(a) minimizes the sum of squared lengths from points to their projections onto any subspace of dimension at most k,

[1]All our projections are orthogonal projections

(b) maximizes the trace of variance matrix of the projected variables onto any subspace of dimension at most k, and
(c) maximizes the sum of squared inter-point distances of the projections onto any subspace of dimension at most k.

PROOF: Without loss of generality we can centre the observations, so each variable has mean zero. Part (a) is follows from the squared Frobenius norm of $X - P_A X$ being the sum of squared lengths of its rows.
For part (b) the squared Frobenius norm of $P_A X$ is the sum of squares of the projected variables, that is $n - 1$ times the sum of the variances of the variables, which is the trace of the variance matrix (and is invariant to the choice of a basis for that subspace).
For (c) consider any projection $P_A X$. Let d_{rs} be the distance between observations r and s, and $\widetilde{d}_{rs}$ the distance under projection (which is smaller, as it *is* a projection). Let $\mathbf{y}_r$ be the rth projected observation as a row vector. Then

$$\sum_{rs} \widetilde{d}^2_{rs} = \sum_{rs} \|\mathbf{y}_r - \mathbf{y}_s\|^2 = \sum_{rs} \|\mathbf{y}_r\|^2 + \|\mathbf{y}_s\|^2 - 2\mathbf{y}_r\mathbf{y}_s^T = 2n \sum_r \|\mathbf{y}_r\|^2 = 2n\|P_A X\|^2$$

which is maximized according to proposition 2. □

2 Principal Components

The traditional definition of principal components is recursive. First choose the linear combination $y = \mathbf{x}a$ of *row* vectors $\mathbf{x}$ of observations which has $\|a\| = 1$ and maximizes the variance of y. Then choose subsequent linear combinations to maximize the variance amongst combinations uncorrelated with those chosen previously. Fix $U\Lambda V^T$ as the SVD of the *centred* data X (that is, with the column means subtracted).

Proposition 4 *The principal components are given, in order, by columns of V. The first k principal components span a subspace with the properties of proposition 3.*

PROOF: Consider a linear combination $y = \mathbf{x}a$ with $\|a\| = 1$. Then

$$\text{var}(y) = a^T \text{var}(\mathbf{x}) a = \frac{1}{n-1} a^T X^T X a = \frac{1}{n-1} a^T V \Lambda V^T a = \frac{1}{n-1} \sum \lambda_i^2 a_i'^2$$

where $a' = V^T a$ also has unit length (and this corresponds to rotating to a new basis of the variables). It is clear that the maximum occurs when a' is the first coordinate vector, or a the first column of V. Now consider the second principal component $\mathbf{x}b$. It must be uncorrelated with the first, so

$$0 = [Xa]^T[Xb] = [U\Lambda a']^T[U\lambda b'] = \lambda_1^2 b_1'$$

and it is obvious that the maximum variance under this constraint is given by taking b' as the second coordinate vector. An inductive argument gives the remaining principal components.

Using the principal component variables, $X = U\Lambda$, so it clear that the subspace spanned by the first k columns is the approximation of propositions 2 and 3. □

The principal components form a useful transformation of the set of variables; they are uncorrelated and have variances $\lambda_1^2/(n-1)$. Thus rescaling the principal components to unit variance 'spheres' the data. On the original variables the variance matrix Σ is given by

$$(n-1)\Sigma = X^T X = V\Lambda^2 V^T$$

so an alternative way to find the principal components is to take the eigendecomposition of Σ; the eigenvalues are then the variances of the principal components. Note that using the correlation matrix rather than the variance matrix is equivalent to re-scaling the original variables to unit variance. Note also that a unit-length combination of principal components has variance in the range of variances of the included principal components, so the last principal component has the smallest variance of any unit-length linear combination.

PROOF: Consider a combination a with $\|a\| = 1$ and $a_1, \ldots, a_{\ell-1} = 0 = a_{r+1}, \ldots, a_p$. Then, on the principal components,

$$\text{var}(\mathbf{x}a) = (n-1)^{-1} a^T \Lambda^2 a = (n-1)^{-1} \sum_{\ell}^{r} a_i^2 \lambda_i^2 \leqslant (n-1)^{-1} \sum_{\ell}^{r} a_i^2 \lambda_\ell^2 = (n-1)^{-1} \lambda_\ell^2$$

and similarly for the lower bound. A unit-length linear combination of the principal components is also a unit-length linear combination of the original variables, by the orthogonality of V. □

Proposition 5 *Consider a orthogonal change XB to k new variables. The first k principal components have maximal variance, both in the sense of the trace and of the determinant of the variance matrix. Similarly, the last k principal components have minimal variance.*

PROOF: The trace statement is proposition 3(b), but we will give an alternative proof. Consider the SVD of XB, and let its singular values be $\mu_1, \ldots, \mu_k$. We will show $\mu_j \leqslant \lambda_j, j = 1, \ldots, k$, which suffices as the trace of the variance matrix is proportional to the sum of the squared singular values, and the determinant is proportional to their product.

Consider a variable $\mathbf{x}a$ which is a unit-length linear combination of the first j principal components of the B set, but is orthogonal to the first $j-1$ original principal components. (A dimension argument shows that such a variable exists. Since B is orthogonal it is also a unit-length combination of the original variables and of their principal components.) This has variance at least μ_j^2 and at most λ_j^2, so $\mu_j \leqslant \lambda_j$.

The result on minimality is proved by showing $\mu_j \geqslant \lambda_{p-k+j}, j = 1, \ldots, k$, taking a unit-length linear combination of the last j original principal components orthogonal to the last $j-1$ principal components of the B set. □

The *Mahalanobis* distance with respect to a covariance matrix Σ between two p-variate row vectors $\mathbf{x}$ and $\mathbf{y}$ is

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})\Sigma^{-1}(\mathbf{x}-\mathbf{y})^T}$$

Note that is the Euclidean distance in the principal component variables re-scaled to unit variance.

3 Metric Scaling

Suppose we choose just to approximate the distances d_{rs} between pairs of observations. Given the distances, we obviously can not recover the observations themselves, since the distances are invariant to rigid motions (including reflections) of $\mathbb{R}^n$. It transpires that that is the only freedom allowed.

Proposition 6 *For any symmetric matrix T, define the matrix*

$$T' = -\frac{1}{2}\left[T - \frac{(T\mathbf{1})\mathbf{1}^T}{n} - \frac{\mathbf{1}(T\mathbf{1})^T}{n} + \frac{\mathbf{1}^T T\mathbf{1}}{n^2}\right]$$

by subtracting row and column means and adding back the overall mean, or, equivalently, by removing row means then column means.

(a) Given any configuration of n points in $\mathbb{R}^p$, the matrix $(d_{rs}^2 = \|\mathbf{x}_r - \mathbf{x}_s\|^2)$ gives a positive semi-definite T'. Such a set of distances is called Euclidean.

(b) Given a symmetric $n \times n$ matrix T with positive semi-definite T', we can find a configuration of points in $\mathbb{R}^{(n-1)}$ such that $T = (d_{rs}^2)$.

(c) A necessary and sufficient condition for a $n \times n$ matrix T to be a squared distance matrix is that $\mathbf{w}^T T\mathbf{w} \leqslant 0$ for all $\mathbf{w}^T\mathbf{1} = 0$.

(d) Any two configurations of n points with the same (d_{rs}^2) differ only by a shift and a rigid motion of $\mathbb{R}^n$, so lie in (shifted) subspaces of the same minimal dimension, the rank of T'.

PROOF: Without loss of generality, centre the data.
(a) $T = (\|\mathbf{x}_r - \mathbf{x}_s\|^2) = (\|\mathbf{x}_r\|^2 + \|\mathbf{x}_s\|^2 - 2\mathbf{x}_r\mathbf{x}_s^T) = E\mathbf{1}^T + \mathbf{1}E^T - 2XX^T$ where $E = (\|\mathbf{x}_r\|^2)$. Let $e = E^T\mathbf{1}$. Then $T\mathbf{1} = nE + e\mathbf{1}$ and $\mathbf{1}^T E\mathbf{1} = 2ne$. Thus

$$-2T' = E\mathbf{1}^T + \mathbf{1}E^T - 2XX^T - E\mathbf{1}^T - e\mathbf{1}\mathbf{1}^T/n - \mathbf{1}E^T - e\mathbf{1}\mathbf{1}^T/n + 2ne\mathbf{1}\mathbf{1}^T/n^2 = -2XX^T$$

which is negative semi-definite.

(b) Let $T' = CD^2C^T$ be the eigendecomposition of T', noting that the eigenvalues are non-negative, and by construction T' has zero column sums and so has rank r at most $(n-1)$. Take X as the first r columns of CD, so $T' = XX^T$. This configuration is centred, since $\|X\mathbf{1}\|^2 = \mathbf{1}^T T'\mathbf{1} = 0$. Note that $(\|\mathbf{x}_r\|^2) = \text{diag}(XX^T) = \text{diag}(T')$, so T' determines $T = (d_{rs}^2)$ and (under zero means) this gives the same T' by result (a).

(c) Note that $[(I - \mathbf{1}\mathbf{1}^T/n)\mathbf{w}]^T T[(I - \mathbf{1}\mathbf{1}^T/n)\mathbf{w}] = -2\mathbf{w}^T T'\mathbf{w}$ which is negative if T' is positive semi-definite.

(d) The procedure of (b) constructs a canonical configuration which is obtained by a shift (to zero mean) and a rigid motion from either configuration. □

Note that since rank$[T']$ = rank$[X - \mathbf{1}(X\mathbf{1})]$, the subspace of (b) is that spanned by the r principal components with $\lambda_i > 0$, and r is the rank of T'.

The claim in Krzanowski (1988) that it is sufficient that the distances satisfy the triangle inequality is incorrect. It does suffice that they are an *ultrametric* (see clustering). [Counter-example due to Dr F.H.C. Marriott: Consider a unit equilateral triangle ABC with centroid O in the plane. We can reduce the distance attributed to AO and keep the triangle inequality. These 4 points if Euclidean must lie in $(R)^3$, and we can take ABC to define a plane. If O lies out of this plane, AO is increased.]

What should we do if the set of distances is not Euclidean? We can seek an approximation by a Euclidean set in $\mathbb{R}^k$ for small k. Note that if the distances *are* Euclidean, the best approximation in $\mathbb{R}^k$ (in the sense of minimizing the difference in squared distances) is the projection onto the first k principal components, by proposition 4, and since $T' = XX^T = U\Lambda^2U^T$, this corresponds to setting $\lambda_{k+1}, \ldots$ to zero. If the distances are not Euclidean, $T' = CDC^T$ is not positive semi-definite, but we can set the negative elements and the small positive elements of D to zero and use the configuration CD.