Risk and Portfolio Management with Econometrics, Courant Institute, Fall 2009
http://www.math.nyu.edu/faculty/goodman/teaching/RPME09/index.html
Jonathan Goodman

Section 3, Singular Value Decomposition (SVD), Principal Component Analysis (PCA)

Let A be an $n \times m$ matrix. Think of column k of A as the column vector a_k. Then A is made up of m of these columns:

$$A = \begin{pmatrix} a_{11} & & a_{1k} & & a_{1m} \\ \vdots & & \vdots & & \vdots \\ a_{j1} & \cdots & a_{jk} & \cdots & a_{jm} \\ \vdots & & \vdots & & \vdots \\ a_{n1} & & a_{nk} & & a_{nm} \end{pmatrix} = \begin{pmatrix} | & & | & & | \\ | & & | & & | \\ a_1 & \cdots & a_k & \cdots & a_m \\ | & & | & & | \\ | & & | & & | \end{pmatrix}$$

so a_k is the column vector

$$a_k = \begin{pmatrix} a_{1k} \\ \vdots \\ a_{jk} \\ \vdots \\ a_{nk} \end{pmatrix}$$

It is useful to keep in mind the common cases $n \gg m$ (tall thin A), and $n \ll m$ (long, short A).

We use the l^2 norm

$$\|a\|_{l^2} = \left(\sum_{k=1}^{n} a_j^2 \right)^{1/2} .$$

In three dimensions, this corresponds to ordinary length. We use no other norm in this section, so we just write it as $\|a\|$. There are other useful norms, such as the l^1 norm of the last homework. But least squares regression and Principal Component Analysis are all l^2 related. That is a strength (there is so much helpful structure in l^2) and a weakness (other norms are more appropriate in many applications).

An $n \times n$ matrix U is *orthogonal* if $U^t U = I$. This is the same as saying that the columns of U form an orthogonal basis of R^n. You should check that the (j,k) entry of $U^t U$ is equal to

$$\left(U^t U \right)_{jk} = u_j^t u_k$$

where the vectors u_k are the columns of U. If $j \neq k$, this says that u_j is perpendicular to u_k. For $j = k$, this says that $\|u_j\| = 1$.

The *Singular Value Decomposition* of A is the decomposition

$$A = U \Sigma V^t , \tag{1}$$

where U is $n \times n$ orthogonal, V is $m \times m$ orthogonal, and Σ is *pseudo-diagonal*, with diagonal entries $(\Sigma_{11} = \sigma_1) \geq (\Sigma_{22} = \sigma_2) \geq \cdots$. The numbers σ_j are the *singular values* of A. Pseudo-diagonal means that $\Sigma_{jk} = 0$ if $j \neq k$. The actual matrix Σ is $n \times m$. If $n > m$, then there are m singular values. There are n if $n < m$. The columns of V are *right singular vectors* and the columns of U are *left* singular vectors. Either the singular vectors of the singular values (or both) are called *principle components*. The matrix Σ does not have to be square and has nothing to do with what we called Σ in earlier sections of notes – the covariance matrix of a multivariate random variable.

In general, a matrix *decomposition* is a factorization of a matrix into the product of other matrices of a specified form. Other decompositions include the QR decomposition: $A = QR$, where Q is orthogonal and R is upper triangular ($R_{jk} = 0$ if $j > k$).

The SVD matrix equation (1) is a compact way of writing relationships involving A and the left and right singular vectors. Multiplying both sides by V, and using $V^t V = I$ on the right, puts (1) in the form $AV = U\Sigma$. In terms of the columns of V and U, this becomes

$$Av_k \;=\; \sigma_k u_k \;, \tag{2}$$

where the v_k are an orthonormal family in R_m and the u_k are an orthonormal family in R^n. If $n > m$, the relations (2) hold only for $k \leq m$. The relations (2) say nothing about the u_k for $k > m$, but we elsewhere have required them to be orthonormal. If $m > n$, then (2) holds only for $k \leq n$. For $k > n$ we have $Av_k = 0$, and also that the v_k form an orthonormal family. Therefore, one thing the SVD does is supply an orthonormal basis of the *kernel* (or *null space*) or A, the set of vectors x with $Ax = 0$.

Taking the transpose of (1) gives $A^t = V^t \Sigma^t U$. Everything we said about columns of A, U, and V has an analogue involving columns of A^t (rows of A), and columns of U^t and V^t with the same singular values σ_k. In particular, if $n > m$, the last $n - m$ columns of U^t (the transposes of the bottom $n - m$ rows of U) form an orthonormal basis of the null space of A^t.

Recall the basic facts about eigenvalues of symmetric matrices. If B is a symmetric $n \times n$ matrix, then B has n real eigenvalues and a family of n orthonormal eigenvectors. The eigenvectors are almost unique if the eigenvectors are distinct. The singular values and singular vectors of A give the eigenvalues and eigenvectors of the matrices AA^t and $A^t A$ (not only are these not equal, one is $m \times m$ and the other is $n \times n$). For example, we have (using $VV^t = I$ at the end)

$$AA^t \;=\; \left(U\Sigma V^t\right)\left(U\Sigma V^t\right)^t \;=\; U\Sigma V^t V \Sigma^t U^t \;=\; U\Lambda U^t \;,$$

Where $\Lambda = \Sigma\Sigma^t$ is an $n \times n$ matrix with the numbers $\lambda_k = \Lambda_{kk} = \sigma_k^2$ on the diagonal. If $n > m$, then we run out of singular values before the end, so $\lambda_k = 0$ for $k > m$. This may be written $AA^t U = U\Lambda$, which in individual columns is

$$\left(AA^t\right) u_k \;=\; \sigma_k^2 u_k \;.$$

This means that the columns of U are the eigenvectors of AA^t, and the eigenvalues AA^t are the squares of the singular values of A. In a similar way, we could see that

$$\left(A^t A\right) v_k \;=\; \sigma_k^2 v_k \;,$$

so the columns of V are the eigenvalues of $A^t A$. Depending on whether $n > m$ or $n < m$, one of the eigenvalues of AA^t or $A^t A$ have to be padded with zeros, once we run out of singular values.

An interesting consequence of this is the **Theorem:** the non-zero eigenvalues of AA^t are equal to the nonzero eigenvalues of $A^t A$. For example, if

$$A \;=\; \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix} \;.$$

then

$$AA^t \;=\; B \;=\; \begin{pmatrix} 3 & 6 \\ 6 & 14 \end{pmatrix} \quad , \quad A^t A \;=\; C \;=\; \begin{pmatrix} 2 & 3 & 4 \\ 3 & 5 & 7 \\ 4 & 7 & 10 \end{pmatrix} \;.$$

According to Matlab, the eigenvalues of B are $\lambda_1 = 16.64$ and $\lambda_2 = .036$. Matlab also says that the eigenvalues of C are $\lambda_1 = 16.64$, $\lambda_2 = .036$, and $\lambda_3 = 0$. Matlab also says that the SVD of A is

$$A \;=\; \begin{pmatrix} -.403 & -.915 \\ -.915 & .403 \end{pmatrix} \cdot \begin{pmatrix} 4.079 & 0 & 0 \\ 0 & .601 & 0 \end{pmatrix} \cdot \begin{pmatrix} -.323 & -.5548 & -.772 \\ -.854 & -.183 & .487 \\ .408 & 0.817 & .408 \end{pmatrix} \;.$$

We can check that $\lambda_1 = \sigma_1^2$, which is $16.64 = 4.08^2$, and $\lambda_2 = \sigma_2^2$, which is $.036 = .601^2$, both of which are true.

1 Orthogonality, variational principles, and the existence theorem

In order to explain the various components of the SVD, here is one of the proofs that the SVD exists. The proof goes by first constructing v_1, u_1, and σ_1, then v_2, u_2, and σ_2, and so on. Some of the properties are obvious and automatic in the construction. In particular, the v_k and u_k will have unit length by definition. The σ_k are positive and decreasing (well, non-increasing and non-negative). The v_k are orthogonal. What is not part of the definition is that the u_k are orthogonal. That comes from the variational principle. Orthogonality often comes from least squares optimization, as happens here.

The starting point is the form for the norm of a matrix

$$\|A\| \;=\; \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} \;. \tag{3}$$

Before discussing what the answer is, we discuss the reason for the answer to exist. Let $f(x)$ be any continuous function of x. We say x_* is a *maximizer* of f over the set B if $f(x_*) \geq f(x)$ for all $x \in B$. If x_* is a maximizer, we say $f(x_*) = \max_{x \in B} f(x)$, and $x_* = \arg\max_{x \in B} f(x)$. A continuous function need not have a maximizer. For example, the function $f(x) = 1/(1+x^2)$ would seem to have maximum value equal to zero, but there is no x_* so that $f(x_*) = 0$.

A theorem in elementary analysis states that the maximizer does exist if f is a continuous function of x and if B is a *compact* set. A set is compact if it is bounded and it contains all its limit points. The maximization problem (3) fails both these tests. It is not bounded because it allows arbitrarily large x. It does not contain the limit point $x = 0$: there is a sequence $x_n \in B$ with $0 = \lim_{n \to \infty} x_n$ not in B. I hope that the theorem seems plausible.

Back to our case (3), we *compactify* the optimization problem by using *scaling*. In particular, for every $x \neq 0$ there is a y with $\|y\| = 1$ and $y = mx$ for some $m > 0$: just take $m = 1/\|x\|$. Note that the factor m does not change the quotient in (3): $\|Ay\| = \|Ay\| / \|y\| = \|Amx\| / \|Amx\| = \|Ax\| / \|x\|$. Therefore, the maximum in (3) is equal to

$$\|A\| \;=\; \max_{\|y\|=1} \|Ay\| \;. \tag{4}$$

The maximum is attained, there is a maximizer, because the function $f(y) = \|Ay\|$ is continuous and the set of allowed values

The norm of a matrix is a single measure of its size. The formula (4) gives $\|A\|$ as the largest amount by which A stretches a vector (by the scaling argument, we need only consider vectors of unit length). We will define $\sigma_1 = \|A\|$. Then σ_2 will be the second largest stretch, σ_3 the third, and so on.

More specifically, let $v_1 = y_*$ in the optimization problem (4) and define σ_1 and[1] u_u by $Av_1 = \sigma_1 u_1$. This makes u_1 a vector of unit length, as it is supposed to be. This is (2) for the case $k = 1$.

We now define σ_2 as the largest stretch possible using a vector perpendicular to v_1:

$$\sigma_2 \;=\; \max_{\|y\|=1,\; v_1^t y = 0} \|Ay\| \;.$$

There is a maximizer, as before, so call it $y_* = v_2$. This satisfies $\|v_2\| = 1$ and $v_1^t v_2 = 0$, as it is supposed to. Define u_2 by $Av_2 = \sigma_2 u_2$, which is (2) with $k = 2$. So the vectors v_1, v_2, u_1, and u_2 satisfy all the properties we asked, except that we do not know that u_1 is perpendicular to u_2. It turns out that this is automatic, which is the main interesting fact that makes the SVD work as it does.

The fact that u_2 is perpendicular to u_1 is a consequence of a simpler fact that is easier to verify:

Lemma. If v_1 is a maximizer of (4) and $v_1^t y = 0$, and $Av_1 = \sigma_1 u_1$, then $u_1^t A y = 0$.

Proof of Lemma. This is a proof by contradiction. We show that if there is a $y \neq 0$ with $v_1^t y = 0$ and $u_u^t A y \neq 0$, then u_1 is not the maximizer of (4). If such

[1] In the lecture, I interchanged u_k and v_k. I hope they are correct here.

a y would exist, we could define a curve in R^n $w(t) = v_1 + ty/\|v_1 + ty\|$. Since y is perpendicular to v_1, the denominator has

$$\frac{d}{dt}\,\|v_1 + ty\|\Big|_{t=0} \;=\; 0\;,$$

so

$$\frac{d}{dt}\,w(t)\Big|_{t=0} \;=\; y\;.$$

The same kind of calculations show that

$$\frac{d}{dt}\,\|Aw(t)\|\Big|_{t=0} \;=\; \sigma_1 u_1^t A y\;,$$

Assuming $\sigma_1 \neq 0$ ($\sigma_1 = 0$ implies that $A = 0$), this shows that the derivative of $\|Aw(t)\|$ is not zero when $t = 0$. The value at $t = 0$ is σ_1. If the derivative is not zero, then σ_1 is not the maximum. This is the contradiction that proves the lemma.

The rest of the singular vectors and singular values are constructed in this way. We find v_3 by maximizing Ay over the set of y with $\|y\| = 1$ and $v_1^t y = 0$ and $v_2^t y = 0$. Then $Av_3 = \sigma_3 u_3$ defines u_3. The orthogonality of u_3 with u_1 and u_2 is not part of the definition, but follows from an argument like the lemma. Clearly $\sigma_1 \geq \sigma_2 \geq \cdots$.

There are three possible endings to this procedure. One is that $m < n$ and we come to $k = m$. Then the m vectors v_k form an orthonormal basis for R^m, but the m vectors u_k are not yet a basis for R^n. We simply find $n - m$ extra orthonormal vectors to make the n vectors u_k into an orthonormal basis of R^n. This provides matrices U and V that work in (1). Another ending is that $m > n$ and we come to $k = n$. We then have constructed n orthonormal vectors u_k that are an orthonormal basis for R^n, but we do not have enough vectors v_k to be a basis for R^m. The argument of the Lemma above shows that any y perpendicular to all the v_k so far has $Ay = 0$. Therefore, we can choose any completing orthonormal vectors to complete the v_k we have into an orthonormal basis. The final possibility is that we come to $\sigma_k = 0$ for $k < n$ and $k < m$. In that case, we complete our existing v_k and u_k in any way.

There are effective computational algorithms for computing the SVD. Running one of them will produce complete orthogonal matrices V and U. As we said above, the last columns of V and U might be somewhat arbitrary.

2 Low rank approximation

The *rank* of a matrix A is the dimension of the space spanned by its columns. It is also the number of non-zero singular values. It happens often in statistics that a most of the non-zero singular values of a large matrix A are very small compared to the largest ones. Setting the small singular values to zero results in

a matrix not much different from A but with much smaller rank. Furthermore, that procedure produces the optimal low rank approximation to A in the least squares sense. This is one of the most important uses of the SVD in practice.

Here are the details. If the columns of A span a space of dimension k, then there are k column vectors that form a basis of that space. Call then w_1, ..., w_k. Since the columns of A are in the space spanned by the w_j, for each i there are numbers b_{ij} so that $a_i = \sum_{j=1}^{k} b_{ij} w_j$. In matrix form, this may be written $A = WB$, where W is the $n \times k$ matrix with columns w_j and B is the $k \times m$ matrix whose entries are the b_{ij}. This shows that a rank k matrix may be written in the form $A = WB$ with the dimensions given. Conversely, if $A = WB$ with those dimensions, then A has rank at most k.

The SVD of A gives a way to construct the W and B above. Suppose that only the first k singular values are different from zero. Then the bottom rows of Σ in (1) are all zeros. This means we get the same result if we use the smaller $k \times k$ matrix $\widetilde{\Sigma}$ that only has the first k singular values (the only non-zero ones) on the diagonal, and only top k rows of V^t and the first k columns of U. The matrix B is $\widetilde{\Sigma}$ multiplying the top k rows of V^t. This is an explicit rank k representation of A.

3 Linear least squares

Least squares problems in statistics are described using different notation from that used above. The object is to estimate the parameters in a linear regression model:

$$y \;=\; b_1 x_1 + \cdots + b_m x_m \;+\; R\,. \tag{5}$$

The idea is that we want to predict y given regressor values $x_1, \ldots, x_m$. More than that, we will do the prediction in a purely linear way. The linear predictor is determined by only of m coefficients $b_1, \ldots, b_m$. The *model* expressed by (**??**) is that R is a Gaussian with mean zero and variance σ^2, where σ^2 is another unknown constant. That is, the observed Y is a linear combination of the x_j combined with a certain Gaussian noise.

The model is *calibrated* using n data rows $(Y_j, X_{j,1}, \ldots, X_{j,m})$. For given coefficients b_i, the *residual* for row i is

$$R_i \;=\; Y_i \;-\; \sum_{j=1}^{m} X_{ij} \widehat{b}_j\,. \tag{6}$$

This is the error in predicting Y_i as a linear combination of the regressors X_{ij} with linear prediction coefficients $\widehat{b}_j$. If we assume that the residuals are independent $\mathcal{N}(0, \sigma^2)$, the maximum likelihood estimator of the b_j results in minimizing the sum of the squares of the residuals:

$$\widehat{b} \;=\; \arg\min_{b} \sum_{i=1}^{n} R_i^2\,. \tag{7}$$

This may be formulated in vector terms. The residual is $R \in R^n$ with entries R_i. The data takes the form of a vector Y and a matrix X. The definition (**??**) then is written as

$$R \;=\; Y \;-\; X\widehat{b}\,, \tag{8}$$

and the goal is to find the vector $\widehat{b} \in R^m$ to minimize

$$\|R\| \;=\; \left\|Y - X\widehat{b}\right\| \;. \tag{9}$$

Of course, minimizing $\|R\|$ is the same as minimizing $\|R\|^2$. One approach to this is the *normal equations* discussed elsewhere.

Here we discuss the approach using the SVD:

$$X \;=\; U\Sigma V^t \;. \tag{10}$$

Substituting (**??**) into (**??**), and using the fact that U and V are orthogonal, gives

$$\left\|Y - U\Sigma V^t\widehat{b}\right\| \;=\; \left\|U^tY - \Sigma V^t\widehat{b}\right\| \;=\; \left\|\widetilde{Y} - \Sigma\widetilde{b}\right\| \;,$$

where

$$\widetilde{Y} \;=\; U^tY \;\;,\quad \widetilde{b} \;=\; V^t\widehat{b}\,.$$

We look in more detail at $\widetilde{R} = \widetilde{Y} - \Sigma\widetilde{b}$ under the assumption that X has full rank m and none of the singular values are zero. It is

$$\begin{pmatrix} \widetilde{R}_1 \\ \vdots \\ \widetilde{R}_m \\ \widetilde{R}_{m+1} \\ \vdots \\ \widetilde{R}_n \end{pmatrix} = \begin{pmatrix} \widetilde{Y}_1 \\ \vdots \\ \widetilde{Y}_m \\ \widetilde{Y}_{m+1} \\ \vdots \\ \widetilde{Y}_n \end{pmatrix} - \begin{pmatrix} \sigma_1 & 0 & \cdots & & 0 \\ 0 & \sigma_2 & 0 & \cdots & \\ \vdots & 0 & \ddots & & 0 \\ & & & \sigma_{m-1} & \\ 0 & \cdots & & 0 & \sigma_m \\ 0 & \cdots & & \cdots & 0 \\ & & & & \\ 0 & & & \cdots & 0 \end{pmatrix} \begin{pmatrix} \widetilde{b}_1 \\ \vdots \\ \widetilde{b}_m \end{pmatrix}$$

In this form, the solution to the least squares problem is obvious. The choice of $\widetilde{b}$ has no effect on the residuals $\widetilde{R}_i$ for $i > m$. On the other hand, choosing

$$\widetilde{b}_i \;=\; \frac{\widetilde{Y}_i}{\sigma_i} \tag{11}$$

sets $\widetilde{R}_i = 0$ for $i \le m$. The choice (**??**) clearly minimizes $\left\|\widetilde{R}\right\|$. Since $\widetilde{R} = U^tR$, this also minimizes $\|R\|$.

The individual component formula (**??**) gives the user detailed control over difficult least squares fitting problems. Many least squares problems are problematic because they are *ill conditioned*. The *condition number* of a least squares problem is

$$\mathrm{cond}(X) \;=\; \frac{\sigma_1}{\sigma_m}\;.$$

This is the ratio of the maximum to minimum stretch. You can think of it as a simple way to measure the range of values in X, in the way $\|X\|$ represents the size of X. Large data sets often have very ill conditioned matrices. If $\sigma_1 \ll \sigma_m$, it is likely that $\widetilde{b}_m \gg \widetilde{b}_1$. As $\sigma_m \to 0$, (**??**) shows that $\widetilde{b}_m \to \infty$. This implies that the $\widehat{b}$, the vector of estimated regression coefficients, has $\left\|\widehat{b}\right\| \to \infty$ as well. A signature of ill conditioned least squares problems is very large regression coefficients.

Regularization means compromising on the exact regression formula (**??**) in order to control the size of the regression coefficients. For example, one common procedures is *Tychanoff regularization*, which replaces (**??**) by

$$\widetilde{b}_i \;=\; \frac{\widetilde{Y}_i}{\left(\sigma_i^2 + \epsilon^2\right)^{1/2}} \tag{12}$$

The denominator is designed so that it is approximately equal to σ_i if σ_i is much larger than the *cutoff* parameter, ϵ. Very small σ_i is (approximately) replaced by ϵ.

The formula (**??**) has a geometric interpretation. Linear least squares regression computes the orthogonal projection of Y onto the columns of X. The residual is what is left over after that projection. The formulas (**??**) accomplish that projection, as we now explain. Suppose first that we have subspace spanned by a single vector, $u \in R^n$, and that $\|u\| = 1$. The projection of Y onto the subspace determined by u is

$$P_u Y \;=\; \left(u^t Y\right) u \; .$$

This is a vector that points in the direction of u. We saw already that the first m columns of U form an orthonormal basis for the subspace of R^n spanned by the columns of X. The formulas (**??**) essentially perform the projection of Y onto the column space of X using the orthonormal basis $u_1, \ldots, u_m$.

4 SVD and low rank approximation

If $x \in R^n$ and U is an $n \times n$ orthogonal matrix, then $\|Ux\| = \|x\|$. You can see this from

$$\|x\|^2 \;=\; \sum_{k=1}^{n} x_k^2 \;=\; x^t x \; .$$

because (transpose reverses the order of the factors, even of one factor is a vector)

$$\|Ux\|^2 \;=\; (Ux)^t Ux \;=\; x^t U^t U x \;=\; x^t x \;=\; \|x\|^2 \; ,$$

Similarly, if B is an $n \times k$ matrix, we can define the *Frobenius* norm in terms of the sums of the squares of the entries:

$$\|B\|_F^2 \;=\; \sum_{ij} b_{ij}^2 \; . \tag{13}$$

This is the same as the sum of the squares of the norms of the columns of B:

$$\|B\|_F^2 \;=\; \sum_{j=1}^{k} \|b_j\|^2 \;.$$

For that reason, we also have

$$\|UB\|_F^2 \;=\; \|B\|_F^2 \;,$$

because $\|Ub_j\|^2 = \|b_j\|^2$, and multiplying B by U is the same as multiplying each column of B by U.

The SVD of X allows us to solve the problem: find the rank k matrix that best approximates X in the least squares sense (**??**):

$$\min_{X_k} \|X - X_k\|_F^2 \;,$$

where the minimization is over all matrices, X_k of rank k. The optimal X_k is easy to find in terms of the SVD of X: use the largest k singular values, together with the corresponding left and right singular vectors.

We see this much as we saw the SVD solution of the linear least squares problem above. Using the fact that U and V are orthogonal matrices, we have (multiply from the left by U^t and from the left by V as before)

$$\|X - X_k\|_F^2 \;=\; \left\|U\Sigma V^t - X_k\right\|_F^2 \;=\; \left\|\Sigma - U^t X_k V\right\|_F^2 \;.$$

This reduces the general optimization problem to the problem

$$\min_{\widetilde{X}_k} \left\|\Sigma - \widetilde{X}_k\right\|_F^2 \;,$$

where Σ is (pseudo-)diagonal, and $\widetilde{X}_k = U^t X_k V$. Not more than ten minutes thought should convince you that the solution to this problem is to take $\widetilde{X}_k = \Sigma_k$, the pseudo-diagonal matrix using the largest k singular values. This gives[2]

$$X_k \;=\; U\Sigma_k V^t \;=\; U_k \Sigma_k V_k^t \;,$$

which uses the largest k singular values and corresponding singular vectors.

This low rank approximation is a form of

[2]There is a slight inconsistency in this formula. In the middle, Σ_k is the $n \times m$ matrix with k non-zero diagonals. On the right, it is the $k \times k$ matrix with the same non-zeros in the same places.