NLA 2021-2022
The singular value decomposition

Martin Sombra

3 November 2021

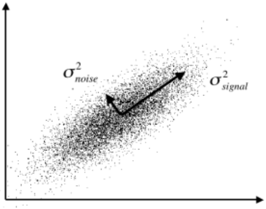
Let M be a data $m \times n$ matrix, for instance:

$$M = \begin{array}{cc} & \begin{array}{cc} \text{age} & \text{height} \end{array} \\ & \left[\begin{array}{cc} \quad & \quad \\ \quad & \quad \\ \quad & \quad \end{array}\right] \text{kids} \end{array}$$

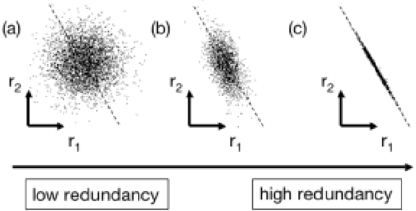
How can we find a simpler description?

Correlated *vs* uncorrelated data

Plotting the samples in $\mathbb{R}^n$ and centering them, we might find that they are *correlated*:



This correlation might be higher *(more significative)* or lower *(less significative)*.



Centering the data

The key parameters in probability and statistics are

$$\boxed{\text{mean}} \quad \text{and} \quad \boxed{\text{variance}}$$

The *mean* of the variables is the row n-vector

$$\mu = \frac{1}{n} \sum_{i=1}^{m} \text{row}_i(M)$$

The *centered data* is the $m \times n$ matrix

$$A = \begin{bmatrix} \text{row}_1(M) - \mu \\ \vdots \\ \text{row}_m(M) - \mu \end{bmatrix}$$

$\rightsquigarrow$ the mean of each variable *(=column)* in A is zero

The covariance matrix

The *covariance matrix* of M is the SPD $n \times n$ matrix

$$S = \frac{1}{m-1} A^T A$$

Its diagonal and off-diagonal entries are the *variances* and *covariances* of the variables in A:

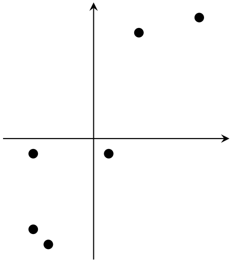
setting $a_j = \mathrm{col}_j(A)$, $j = 1, \ldots, n$, we have that

$$s_{k,k} = \frac{\langle a_k, a_k \rangle}{m-1} = \mathrm{var}(a_k) \quad \text{and} \quad s_{k,l} = \frac{\langle a_k, a_l \rangle}{m-1} = \mathrm{cov}(a_k, a_l)$$

Example

Consider the centered data matrix of ages and weights

$$A = \begin{bmatrix} 3 & 7 \\ -4 & -6 \\ 7 & 8 \\ 1 & -1 \\ -4 & -1 \\ -3 & -7 \end{bmatrix}$$



Its covariance matrix is

$$S = \frac{1}{6-1} A^T A = \begin{bmatrix} 20 & 25 \\ 25 & 40 \end{bmatrix}$$

The total variance

The *total variance* of the variables in A is

$$\text{var}(A) = \sum_{j=1}^{n} \mathrm{var}(a_j) = \frac{1}{m-1} \sum_{i,j} a_{i,j}^2 = \frac{1}{m-1} \|A\|_{\mathrm{F}}$$

Consider the full SVD

$$A = U \Sigma V^T$$

By the orthogonal invariance of the Frobenius norm:

$$\text{var}(A) = \frac{1}{m-1} \|U \Sigma V^T\|_{\mathrm{F}} = \frac{1}{m-1} \|\Sigma\|_{\mathrm{F}} = \frac{1}{m-1} \sum_{j=1}^{n} \sigma_j^2$$

Orthonormal changes of variables

Let q_i, $i = 1, \ldots, n$, be an orthonormal basis of $\mathbb{R}^n$

For an n-vector x, its representation with respect to this basis is

$$x = \sum_{j=1}^{n} \langle q_j, x \rangle \, q_j = \sum_{j=1}^{n} (x^T q_j) \, q_j$$

Consider the orthogonal $n \times n$ matrix $Q = [q_1 \cdots q_n]$ and set

$$B = A\,Q = \begin{bmatrix} b_1 & \cdots & b_n \end{bmatrix}$$

so that $b_{i,j} = \mathrm{row}_i(A)\, q_j$

We have that

- the j-variable in B is the linear combination $\langle q_j, x \rangle$ of the variables in A
- for each k, the $m \times k$ matrix $B_k = \begin{bmatrix} b_1 & \cdots & b_k \end{bmatrix}$ gives the projection of the centered data matrix A into the k-th linear subspace

$$\mathrm{span}(b_1, \ldots, b_k)$$

Orthonormal changes of variables (cont.)

By the orthogonal invariance of the Frobenius norm

$$\mathrm{var}(B) = \frac{1}{m-1}\|B\|_{\mathrm{F}} = \frac{1}{m-1}\|A\|_{\mathrm{F}} = \mathrm{var}(A)$$

$$\boxed{\text{the total variance is invariant by orthogonal changes of variables}}$$

Consider again the full SVD

$$A = U\,\Sigma\,V^T$$

For $Q = V$ we have that

$$\mathrm{var}(B_k) = \mathrm{var}(A\,V_k) = \frac{1}{m-1}\|U\,\Sigma_k\|_{\mathrm{F}} = \frac{1}{m-1}\sum_{j=1}^{k}\sigma_j^2$$

This is the *maximal* total variance among all possible orthogonal projections of the data into a k-th linear subspace of $\mathbb{R}^n$

The smallest distance

Also, the sum of the squares of the distances between the samples and its projections is *minimal* for this k-th linear subspace:

$$
\begin{aligned}
\sum_{i=1}^{m} \left\| \mathrm{row}_i(A) - \sum_{j=1}^{k} b_{i,j}\, v_j \right\|_2^2 &= \sum_{i=1}^{m} \left\| \sum_{j=k+1}^{n} b_{i,j}\, v_j \right\|_2^2 \\
&= \| A \cdot [0 \ \cdots \ 0 \ v_{k+1} \ \cdots \ v_n] \|_{\mathrm{F}} \\
&= \sum_{j=k+1}^{n} \sigma_j^2
\end{aligned}
$$

This is a consequence of the Eckart-Young theorem

Principal components and principal directions

The *j-th principal direction* and the j-th *principal component* of the data matrix M are

$$q_j \in \mathbb{R}^n \quad \text{and} \quad b_j = \begin{bmatrix} \text{row}_1(A)\, q_j \\ \vdots \\ \text{row}_m(A)\, q_j \end{bmatrix} \in \mathbb{R}^m$$

These variables are not correlated, and are ordered according to their variances:

$$\langle b_i, b_j \rangle = 0 \text{ for all } i \neq j \quad \text{and} \quad \langle b_i, b_i \rangle = \sigma_i^2 \text{ for each } i$$

Principal components and principal directions (cont.)

The principal components $b_i = \sigma_i\, u_i$, $i = 1, \ldots, k$, account for

$$\frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{n} \sigma_i^2}$$

of the total variance of the data

A good choice of k keeps the true *signal* and discards the *noise*

Example

The full SVD of A is given by

$$\begin{bmatrix} 3 & 7 \\ -4 & -6 \\ 7 & 8 \\ 1 & -1 \\ -4 & -1 \\ -3 & -7 \end{bmatrix} = \begin{bmatrix} -0.44 & -0.36 \\ 0.43 & 0.01 \\ -0.63 & 0.33 \\ 0.02 & 0.35 \\ 0.18 & -0.70 \\ 0.44 & 0.37 \end{bmatrix} \begin{bmatrix} 16.87 & 0 \\ 0 & 3.92 \end{bmatrix} \begin{bmatrix} -0.56 & -0.83 \\ 0.83 & -0.56 \end{bmatrix}$$

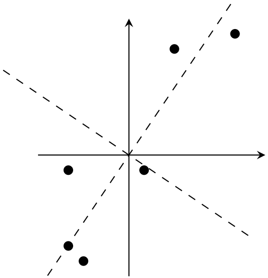
- The columns of $U\Sigma$ are the first and second principal components, and the columns of V give the first and second principal directions
- The variances and covariances of the principal components are given by

$$\frac{1}{6-1}(AV)^T(AV) = \begin{bmatrix} 56.92 & 0 \\ 0 & 3.07 \end{bmatrix},$$

Example (cont.)

Graphically

$$\begin{bmatrix} 3 & 7 \\ -4 & -6 \\ 7 & 8 \\ 1 & -1 \\ -4 & -1 \\ -3 & -7 \end{bmatrix} = \begin{bmatrix} -0.44 & -0.36 \\ 0.43 & 0.01 \\ -0.63 & 0.33 \\ 0.02 & 0.35 \\ 0.18 & -0.70 \\ 0.44 & 0.37 \end{bmatrix} \begin{bmatrix} 16.87 & 0 \\ 0 & 3.92 \end{bmatrix} \begin{bmatrix} -0.56 & -0.83 \\ 0.83 & -0.56 \end{bmatrix}$$



The rank deficient LSP revisited

The SVD also applies to the LSP, and it is particularly appropriate in the rank deficient case

Let A be an $m \times n$ matrix of rank r. If $r < n$ then the solution $x_{\min}$ of the LSP

$$\min_{x \in \mathbb{R}^n} \|A\,x - b\|_2$$

is not unique since

$$A\,x_{\min} = A\,(x_{\min} + y)$$

for any $y \in \mathrm{Ker}(A) \simeq \mathbb{R}^{n-r}$

A reasonable choice is the minimizer having the *smallest norm*

The full and the reduced SVD's

$$\begin{array}{cc} & n \\ m & \begin{bmatrix} \; A \; \end{bmatrix} \end{array} = \begin{array}{cc} & m \\ m & \begin{bmatrix} \; U_r \; \end{bmatrix} \end{array} [\Sigma][V]^T = \begin{array}{cc} & r \\ m & \begin{bmatrix} \; U_r \; \end{bmatrix} \end{array} [\Sigma_r][V_r]^T,$$

that is, if $U = [u_1 \ \cdots \ u_m]$ and $V = [v_1 \ \cdots \ v_n]$ then $U_r = [u_1 \ \cdots \ u_r]$ and $V_r = [v_1 \ \cdots \ v_r]$, and if

$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & \mathbb{0} & \end{bmatrix}$$

then $\Sigma_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$

The Moore-Penrose inverse and the LSP

The *Moore-Penrose inverse* (or *pseudo-inverse*) of A is the $n \times m$ matrix

$$A^+ = V_r \Sigma_r^{-1} U_r^T$$

$\rightsquigarrow$ the solution of the LSP can be written as

$$x_{\text{min}} = A^+ b$$

When A is rank deficient, it is the solution with the *smallest norm*

The Moore-Penrose inverse and the LSP (cont.)

The solution given by the pseudo-inverse is well-conditionned when the smallest nonzero singular value of A is not too small:

Changing b to $b + \delta b$ changes x to $x + \delta x$ with

$$\|\delta x\|_2 \leqslant \frac{\|\delta b\|_2}{\sigma_r}$$

because

$$\delta x = V_r \Sigma_r^{-1} U_r^T \delta b$$

and so

$$\|\delta x\|_2 \leqslant \|\Sigma_r^{-1}\|_2 \, \|\delta b\|_2 = \frac{\|\delta b\|_2}{\sigma_r}$$

Example

Let

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Then

$$A^+ = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

and so

$$\widehat{x} = A^+ b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

with condition number $\frac{1}{\sigma_1} = 1$

Setting $A_\varepsilon = \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon \end{bmatrix}$ for $\varepsilon > 0$ gives

$$\widehat{x}_\varepsilon = \begin{bmatrix} 1 \\ \frac{1}{\varepsilon} \end{bmatrix}$$

A practical strategy

The rank is not continuous, and so it might be affected by small perturbations: in the example, round off will make perturbations of size

$$O(\varepsilon)\,\|A\|_2$$

that might increase the condition number from $\frac{1}{\sigma_r}$ to $\frac{1}{\varepsilon}$

The SVD is *backward stable*: round off gives

$$(U+\delta U)\,(\Sigma+\delta\Sigma)\,(V+\delta V)^T = A+\delta A$$

with $\|\delta A\|_2 \leqslant O(\varepsilon)\,\|A\|_2$

$\rightsquigarrow$ the computed singular values $\sigma_i + \delta\sigma_i$ verify

$$|\delta\sigma_i| \leqslant O(\varepsilon)\,\|A\|_2$$

A practical strategy

Let $\mathrm{tol} > 0$ be a user supplied measure of uncertainty in A, e.g.

$$\mathrm{tol} = C\,\varepsilon\,\|A\|_2$$

for some small constant $C > 0$

Given the computed factors in the SVD of A

$$\widetilde{U}, \quad \widetilde{\Sigma}, \quad \widetilde{V}$$

set

$$\widehat{\sigma}_i = \begin{cases} \widetilde{\sigma}_i & \text{if } \widetilde{\sigma}_i \geqslant \mathrm{tol} \\ 0 & \text{else} \end{cases}$$

A practical strategy

Replace $\widetilde{\Sigma}$ by the *truncated SVD*

$$\widehat{\Sigma} = \begin{bmatrix} \widetilde{\sigma}_1 & & & & & \\ & \ddots & & & & \\ & & \widetilde{\sigma}_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \\ & & & \mathbb{0} & & \end{bmatrix}$$

Then setting

$$\widehat{x} = \widetilde{V}_r \widehat{\Sigma}_r^{-1} \widetilde{U}_r^T$$

the error is bounded by $O(\mathrm{tol})$ and the condition number by $\frac{1}{\sigma_r}$