Applied Multivariate Techniques

Daniele Zago

January 21, 2022

Contents

Lecture 1: Principal component analysis

2022-01-13

1.1 Matrix algebra review

We usually have $X_{n\times p} = \begin{pmatrix} x_{11}, \dots, x_{ip} \\ \\ x_{n1}, \dots, x_{np} \end{pmatrix} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix}$.

Def. (Orthogonal matrix)
A square matrix Q is ***orthogonal*** if $Q^\top Q = I$.

Properties

- $Q^{-1} = Q^\top$
- $QQ^\top = I$
- $|Q| = \pm 1$
- A, B orthogonal matrices, then $A^\top B = C$ is still an orthogonal matrix.

  Proof.
  $C^\top C = B^\top A A^\top B = B^\top B = I$.

□

Def. (Eigenvalue)
Let A be a $p \times p$ square matrix, then the roots to the characteristic polynomial

$$q(\lambda) = \det(A - \lambda I)$$

$\{\lambda_1, \dots, \lambda_p\} \in \mathbb{C}$, are called ***eigenvalues*** .

Def. (Eigenvector)
For each λ_i eigenvalue, there exists a unique ***eigenvector*** γ_i associated to λ_i such that

$$A\gamma_i = \lambda_i \gamma_i.$$

The uniqueness is constrained to $\gamma_i^\top \gamma_i = 1$.

Remark If A is symmetric, then $\lambda_i \in \mathbb{R}$ for all i.

Remark If all $\lambda_i > 0$ we have that $x^\top A x > 0$ for any $x \in \mathbb{R}^p$ and A is called ***positive-definite***. If all $\lambda_i \geq 0$, then $x^\top A x \geq 0$ and A is ***positive-semidefined***.

Def. (Rank)
The ***rank*** of A is defined as $\operatorname{rank} A = \#(\lambda_i > 0)$.

Properties If λ_i are eigenvalues for A, then

1. *Addition*: $|A + \alpha I - (\lambda + \alpha)I| = 0 \implies \alpha + \lambda_i$ are eigenvalues of $A + \alpha I$.
2. *Multiplication*: $|\alpha A - \alpha\lambda I| = 0 \implies \alpha \cdot \lambda_i$ are eigenvalues of αA.

Theorem 1 (Spectral decomposition)
A symmetric matrix A can be written in terms of its eigenvalues and eigenvectors as

$$A_{p\times p} = \Gamma\Lambda\Gamma^\top = \sum_{i=1}^{p} \lambda_i \gamma_i \gamma_i^\top,$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ and $\Gamma = (\gamma_1\ \gamma_2\ \ldots\ \gamma_p)$ is orthonormal, where eigenvalues and eigenvectors are counted with multiplicity and $\lambda_1 > \lambda_2 > \ldots > \lambda_p$.

Power With the above decomposition, we have a fast way of computing A^q,

$$A^q = \Gamma\Lambda^q\Lambda^\top.$$

If $A \succ 0$ then $q \in \mathbb{Q} \setminus 0$, else if $A \succeq 0$, then $q \in \mathbb{Q}^+$.

Principal components We have that γ_1 is the solution to the following maximization problem

$$\gamma_1 = \underset{x}{\text{argmax}}\, x^\top Ax \implies \gamma_1^\top A\gamma_1 = \gamma_1^\top\Gamma\Lambda\Gamma^\top\gamma_1 = \lambda_1.$$

The second eigenvalue maximizes

$$\gamma_2 = \underset{\substack{x^\top x=1 \\ x^\top\gamma_1=0}}{\text{argmax}}\, x^\top Ax$$

1.2 Singular value decomposition

Theorem 2 (Singular value decomposition)
Let $X_{n\times p}$ be a general matrix, then X can be written as

$$X_{n\times p} = U_{n\times n}D_{n\times p}V'_{p\times p} = \sum_{i=1}^{\min\{n,p\}} d_i u_j v_i^\top,$$

where $UU^\top = I_n$, $VV^\top = I_p$, and

$$D = \begin{pmatrix} d_1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & d_2 & 0 & 0 & \ldots & 0 \\ 0 & 0 & d_3 & 0 & \ldots & 0 \\ 0 & 0 & 0 & d_4 & \ldots & 0 \\ 0 & 0 & 0 & 0 & \ldots & d_{\min\{n,p\}} \\ 0 & 0 & 0 & 0 & \ldots & 0 \\ \vdots & & & & & \end{pmatrix}$$

Remark The matrix D has lots of zeros, therefore if we set $d_i = 0$ for $i \geq h$, effectively truncating the approximation to the first h components, we obtained a compressed representation of X.

Example (Linear model)
Consider a linear model $y = X\beta + \varepsilon$, and let P be the projection matrix on the parameter model estimates, i.e.

$$Py = X(X^\top X)^{-1}X^\top y,$$

and consider now the singular value decomposition of $X = UDV^\top$:

$$\begin{aligned} P &= UDV^\top(VDU^\top UDV^\top)^{-1}VDU^\top \\ &= UDV^\top(VD^2V)^{-1}VDU^\top \\ &= UDV^\top VD^{-2}VVDU^\top \\ &= UDD^{-2}DU^\top \\ &= UU^\top. \end{aligned}$$

We have that $U_{n\times p}$ is a semi-orthogonal matrix, therefore $U^\top U = I_p$ but $UU^\top \neq I_n$.

Suppose now that we apply a linear transformation on X before computing the estimates, i.e.

$$Z = XC, \quad C \text{ orthogonal and } \operatorname{rank} C = p,$$

then $Z = UD(VC)^\top = UDC^\top V^\top$ and the projection matrix P_Z can be calculated as

$$\begin{aligned} P_Z &= UDC^\top V^\top(VCDU^\top UDC^\top V^\top)^{-1}VCDU^\top \\ &= UDC^\top V^\top VCD^{-2}C^\top V^\top VCDU^\top \\ &= UU^\top \end{aligned}$$

Remark The above result is slightly more complicated but still holds if C is not orthogonal but has rank p.

Exercise: Let P be the projection matrix of rank p, then prove that the eigenvalues are all 1 and that the ***residual maker matrix*** $I - P$ has rank $n - p$ and $\lambda_1, \ldots, \lambda_{n-p} = 1$.
Use the properties of the eigenvalues (addition/multiplication).

Def. (Centering matrix)
Consider the matrix $H = \frac{1}{n}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix} = \mathbb{1}(\mathbb{1}^\top\mathbb{1})^{-1}\mathbb{1}^\top$, which is the linear model when only the intercept term is available. Then, $I - H$ is the ***centering matrix*** and

$$X_C := (I - H)X,$$

which has column-wise zero mean.

Notation Starting from now, we will not use a centering matrix anymore, and we assume that X has been already centered.

Def. (Variance-covariance matrix)
For a centered matrix X we define $\frac{1}{n}X^\top X$ as the ***variance-covariance*** matrix of X.

We want to look at the first principal component, which is defined as the direction g such that

$$\begin{aligned}\widehat{g} &= \underset{g}{\operatorname{argmax}}\, \mathbb{V}[Xg] \\ &= \underset{g}{\operatorname{argmax}}\, g^\top X^\top X g,\end{aligned}$$

and this problem has the solution given by the singular value decomposition of $X^\top X$. Let

$$X^\top X \overset{\text{s.d.}}{=} \Gamma\Lambda\Gamma^\top,$$

we know that the maximum is attained in the first eigenvector, $\widehat{g} = \gamma_1$.

Def. (Principal components)
The j^{th} principal component is the j^{th} direction of maximum variance constrained to being uncorrelated with the previous $j-1$ directions of maximum variance, and

$$g_j = \gamma_j,$$

where γ_j are the eigenvectors of the SVD of $X^\top X$.

The variance of the i^{th} principal components are given by

$$X\gamma_i = \frac{1}{n}\lambda_i.$$

The fraction of explained variance by the i^{th} principal component is

$$\%\mathbb{V}_i = \frac{\lambda_i/n}{\operatorname{tr}(\Lambda)/n} = \frac{\lambda_i}{\sum_{i=1}^p \lambda_i}.$$

We now describe the connection between the SVD and X and the SVD of $X^\top X$:

$$X^\top X \xrightarrow{\text{sp. dec.}} \Gamma\Lambda\Gamma^\top$$

$$X \xrightarrow{\text{sing. val.}} UDV^\top$$

Then, we have that

$$X^\top X = VDU^\top UDV^\top = VD^2V,$$

therefore the singular value decomposition of X is such that $V = \Gamma$ and $D = \Lambda^2$.

In general, the principal components of X are defined by UD, therefore we can obtain them by applying the transformation

$$UD = X\Gamma.$$

Example (Problems with SVD)
Suppose we have a biometric test where we have different unit of scale: if we change the unit of measurement, we get different results in terms of principal components.

To do so, we usually apply the SVD to the standardized variables whenever we do not have variables on the same scale.

Exercise Prove that $P = X(X^\top X)^{-1}X^\top$ has rank p, then prove that $(I - P)$ has rank $n - p$ and find the possible eigenvalues of P and $I - P$.

Proof.
Since P is the projection matrix on $\langle x_1, x_2, \dots, x_p\rangle$, we have that

$$\operatorname{rank} P = \operatorname{rank} X = p.$$

Moreover, we have that $I - P$ is the projection matrix on the orthogonal subspace $\langle x_1, x_2, \dots, x_p\rangle^\perp$, which is a linear subspace of dimension $n - p$, and therefore $\operatorname{rank} I - P = n - p$.

Since the projection matrix P is such that $P = P^2$, we have that if λ is an eigenvalue of P relative to an eigenvector v, then

$$\lambda^2 v = P^2 v = Pv = \lambda v,$$

hence $\lambda^2 = \lambda$, and this can only happen if either $\lambda = 0$ or $\lambda = 1$. The same applies for $(I - P)$, since $(I - P)^2 = (I - P)$.

□

Exercise Let $X_{n\times p}$ and $P_X = X(X^\top X)^{-1}X^\top$ be the projection matrix, let now R be a rotation matrix such that $R^\top R = I$ and $RR^\top = I$. Define $Y = XR$, prove that $P_Y = Y(Y^\top Y)^{-1}Y^\top = P_X$.

Proof.

$$\begin{aligned}
P_Y &= Y(Y^\top Y)^{-1}Y^\top \\
&= XR^\top(R^\top X^\top XR)^{-1}R^\top X^\top \qquad (\text{since } R^{-1} = R^\top) \\
&= XR^\top R^\top (X^\top X)^{-1}RR^\top X^\top \\
&= X(X^\top X)^{-1}X^\top.
\end{aligned}$$

□

Lecture 2: Multidimensional Scaling

2022-01-20

MultiDimensional Scaling (MDS) is a technique which starts from an observed matrix D of pairwise distances and aims to reconstruct an approximate low-dimensional ***configuration*** of points which could have produced D. This in turn is very useful for obtaining a low-dimensional representation of the data in order to visualize clusters and extract relevant information.

Suppose that we have $x_1, \dots, x_n$ observations in a general space $\mathbb{R}^p$, and we know the distances between each pair of elements $d_{ij} = d(x_i, x_j)$. This distance can be any arbitrary distance function, as long as it satisfies the three following properties

1. $d_{ij} \geq 0$ and $d_{ij} = 0 \iff i = j$.
2. $d_{ij} = d_{ji}$
3. $d_{ij} + d_{jk} \geq d_{ik}$

Example (Euclidean distance in $\mathbb{R}^p$)
If $x_i, \dots, x_n \in \mathbb{R}^p$, then

$$d_{ij} = \sqrt{(x_i - x_j)^\top (x_i - x_j)} = \|x_i - x_j\|_2.$$

Note Since we can start from an arbitrary distance matrix D, it's possible to apply the multidimensional scaling even without knowing a) the original data which produces D and b) the true dimension of the underlying space.

Def. (Multidimensional scaling)
Consider the observed symmetric square matrix of distances $D_{n\times n} = (d_{ij})_{i,j=1,\dots,n}$, the ***multidimensional scaling*** (MDS) procedure aims to obtain a low-dimensional representation $z_1, \dots, z_n \in \mathbb{R}^k$ such that

$$z_1, \dots, z_n = \underset{v_1,\dots,v_n}{\operatorname{argmin}} \sum_{i,j} (d_{ij} - \|v_i - v_j\|_2)^2. \tag{1}$$

Interpretation With the above minimization, we obtain a low-dimensional representation of the higher-dimensional observed data. The obtained configuration is thus as similar as possible in terms of distance structure to the original points $x_1, \dots, x_n$.

Notation

- We denote by $D_2 = (d_{ij}^2)_{i,j}$ the matrix of squared distances, and note that $D_2 \neq D^2$.
- We also define the residualizing matrix by $H = I - \frac{1}{n}\mathbb{1}\mathbb{1}^\top = I - \frac{1}{n}J$

› Finally, we define $B = -\frac{1}{2} H D_2 H$, which is a ***double-centering*** of D_2. The resulting row-wise and column-wise sums are both zeros:

$$\mathbb{1}^\top H D_2 H = \mathbf{0}$$

$$H D_2 H \mathbb{1} = \mathbf{0}^\top$$

There is a very strong connection between the principal component analysis and the multidimensional scaling.

Def. (Euclidean matrix)
We say that the matrix $D = (d_{ij})_{i,j}$ is ***euclidean*** if there exists a configuration $z_1, \dots, z_n \in \mathbb{R}^p$ such that $d_{ij} = \|z_i - z_j\|_2$

Note In the following, we denote by Z the matrix of the corresponding configuration of n vectors,

$$Z = \begin{pmatrix} z_1^\top \\ z_2^\top \\ \vdots \\ z_n^\top \end{pmatrix}. \tag{2}$$

Theorem 3 (Euclidean matrix and B matrix)
Let D be a matrix and define $B = -\frac{1}{2} H D H$, then D is euclidean $\iff$ B is positive semidefinite. We call the matrix B the ***inner product matrix****.*

Proof.
Since $-2B = H D_2 H$ and $H = I - \frac{1}{n} J$, then

$$\begin{aligned} -2B &= D_2 H - \frac{1}{n} J D_2 H \\ &= D_2 - \frac{1}{nJ} - \frac{1}{n} J D_2 + \frac{1}{n} J D_2 J. \end{aligned}$$

For each element of $-2B$, we have

$$(-2B)_{ij} = d_{ij}^2 - \frac{1}{n} \sum_h d_{ih}^2 - \frac{1}{n} \sum_k d_{kj}^2 + \sum_h \sum_k \frac{1}{n^2} d_{hk}, \tag{3}$$

now since D is euclidean, we can express d_{ij} in terms of a distance between each element z_i and z_j, $d_{ij}^2 = (z_i - z_j)^\top (z_i - z_j) = z_i^2 - 2z_i z_j + z_j^2$, hence

$$\frac{1}{n}\sum_h d_{ih}^2 = \frac{1}{n} n z_i^2 + \sum_h \frac{z_h^2}{n} - 2z_i \frac{1}{n}\sum_h z_h$$

$$= z_i^2 + \sum_{i=1}^{h} \frac{z_h^2}{n} - 2z_i \bar{z}$$

$$\frac{1}{n}\sum_{h,k} d_{hk} = \frac{\sum_h z_h^2}{n} + \frac{\sum_h z_h^2}{n} - 2\bar{z}^2$$

If we substitute the above terms in Equation (3), then we obtain (exercise)

$$(-2B)_{ij} = -2(z_i - \bar{z})^\top (z_j - \bar{z}).$$

□

Note We have that $B = (b_{ij})_{i,j} = (z_i^\top z_j)_{i,j}$, hence the name *inner product matrix*,

$$B = HZ(HZ)^\top.$$

2.1 Relationship with PCA

The following theorem states the link between the metrix MDS and the principal component, and gives an algorithm for immediately obtaining the solution to the MDS problem (1).

Theorem 4 (MDS and principal components)
Let D be euclidean, then if we define Z as (2), we have that if $B \geq 0$, then there exists $Z = US$ such that

$$B = US^2U^\top,$$

where $UU^\top = I$ and $S^2 = \mathrm{diag}(s_1^2, s_2^2, \ldots, s_k^2)$

Remark From the above theorem, if we compute the singular value decomposition on a positive-semidefined $B = -\frac{1}{2}HD_2H$, then we obtain a representation $Z = US$ which minimizes the multidimensional scaling problem.

Low-dimension If we choose a lower-dimensional representation, say $z_1, \ldots, z_n \in \mathbb{R}^k$ with $k < p$, then we obtain the *optimal* configuration with minimal discrepancy from the observed matrix D.

Proof.
Define $B = US^2U^\top = ZZ^\top$, where $Z = US$. Then, we know that if we write

$$(z_i - z_j)^\top (z_i - z_j) = z_i^2 + z_j^2 - 2z_i z_j$$

$$= b_{ii} + b_{jj} - 2b_{ij},$$

but then we can write each b_{ij} in terms of the distances, since $B = -\frac{1}{2}HD_2H$. Check that

$$\begin{aligned} b_{ij} &= d_{ij}^2 - \frac{1}{n}\sum_h d_{ih}^2 - \frac{1}{n}\sum_k d_{kj}^2 + \frac{1}{n^2}\sum_{h,k} d_{hk}^2 \\ &= -\frac{1}{2}(-2d_{ij}^2) \\ &= d_{ij}^2. \end{aligned}$$

□

2.2 Non-metric MDS

The above discussion states the optimality of metric MDS, i.e. when D is euclidean, and its equivalence to principal component analysis. However, most of the times D is not euclidean and the resulting matrix $B = -\frac{1}{2}HD_2H$ is not guaranteed to have non-negative eigenvalues. Therefore, a lot of research has developed non-metric variants of the multidimensional scaling procedure, which extend its analysis to more general ***dissimilarity metrics***.

Exercise On Moodle, try to analyze the uploaded dataset using the MDS approach.