Canonical Correlation Analysis

Uri Shaham

October 11, 2022

1 Rayleigh quotient

Definition 1.1. *Let $A \in \mathbb{R}^{m \times m}$ be a symmetric matrix, and let $x \in \mathbb{R}^m$ be a nonzero vector. Their Rayleigh quotient is defined as*

$$R(A,x) = \frac{x^T M x}{x^T x}. \tag{1}$$

Theorem 1.2. $\max_x R(A,x) = \lambda_1(A)$ *and the corresponding vector $v_1(A)$ is the maximizer.*

Proof. Let $A = V\Lambda V^T$ be the eigendecomposition of A. Define $y = V^T x$. Then

$$\begin{aligned} R(A,x) &= \frac{x^T A x}{x^T x} \\ &= \frac{x^T V \Lambda V^T x}{x^T V V^T x} \\ &= \frac{y^T \Lambda y}{y^T y}. \end{aligned} \tag{2}$$

Hence, to maximize the R(A, x) one needs to find a unit vector y which maximizes $u^T \Lambda y$.

$$y^T \Lambda y = \sum_{i=1}^m y_i^2 \lambda_i(A) \leq \lambda_1(A) \sum_{i=1}^m y_i^2 = \lambda_1(A).$$

Observe that $y = (1, 0, \ldots, 0)$ gives $y^T \Lambda y = \lambda_1(A)$. Hence $x = Vy = v_1(A)$. $\square$

2 Principal component analysis

Let $x \in \mathbb{R}^m$ be a random vector with zero mean. in PCA we seek for a linear combination of $(x_1, \ldots, x_m)$ along which the variance is maximized. This is the first principal direction. for each subsequent direction, we seek to maximize the variance and in the orthogonal complement of the subspace of previously selected components.

Formally, to find the first direction, w, we maximize

$$\mathrm{Var}\left(\frac{w^T}{\|w\|} x\right) = \frac{w^T \mathbb{E}\left[x x^T\right] w}{w^T w}. \tag{3}$$

Equation (3) is a Raighley quotient and hence its maximizer is the vector $w \in \mathbb{R}^m$ which is the eigenvector of the covariance matrix $\mathrm{Cov}(x) = \mathbb{E}\left[x x^T\right]$ with the largest eigenvalue.

Following the same logic, and since all principal are orthogonal, subsequent directions are the next eigenvectors. Let C be the sample covariance matrix, computed a matrix $X \in \mathbb{R}^{n\times m}$ of from n i.i.d samples of x^T (i.e., $C \in \mathbb{R}^{m\times m}$. Let $C = V\Lambda V^T$ be the eigendecomposition. The projection of the data onto the principal directions is then $X^T V$. To map the data back to the original coordinates one simply multiply the projected data matrix from the right be V^T.

3 Canonical correlation analysis

Let $x \in \mathbb{R}^n,\ y \in \mathbb{R}^m$ be random vectors. In CCA we seek for linear combinations of $(x_1, \dots, x_n)$ and $(y_1, \dots, x_m)$ which are maximally correlated. Put another way, this finds bases for x and y such that the correlation matrix in the new bases is diagonal and correlations on the diagonal are maximized.

Let's start with finding the first pair of projections, $u_1 = a^T x, v_1 = y^t b$. Assume that both x and y are zero mean. Then

$$\begin{aligned}
\rho_1 &:= \operatorname{corr}(u_1, v_1) \\
&= \frac{\mathbb{E}\left[u_1 v_1\right]}{\sqrt{\mathbb{E}\left[u_1^2\right]}\sqrt{\mathbb{E}\left[v_1^2\right]}} \\
&= \frac{\mathbb{E}\left[a^T x y^T b\right]}{\sqrt{\mathbb{E}\left[A^T x x^T a\right]}\sqrt{\mathbb{E}\left[b^T y y^T b\right]}} \\
&= \frac{a^T \Sigma_{x,y} b}{\sqrt{a^T \Sigma_{xx} a}\sqrt{b^T \Sigma_{yy} b}},
\end{aligned}$$

where $\Sigma_{xx}, \Sigma_{yy}, \Sigma_{xy}$ are the covariance and cross-covariance matrices.

Next, we change the basis and define $c = \Sigma_{xx}^{\frac{1}{2}} a,\ c = \Sigma_{yy}^{\frac{1}{2}} b$. Then

$$\rho_1 = \frac{c^T \Sigma_{xx}^{-\frac{1}{2}} \Sigma_{xy} \Sigma_{yy}^{-\frac{1}{2}} d}{\sqrt{c^T c}\sqrt{d^T d}}. \tag{4}$$

Applying Cauchy-Schwartz inequality on the numerator of equation (4) we have

$$\left(c^T \Sigma_{xx}^{-\frac{1}{2}} \Sigma_{xy} \Sigma_{yy}^{-\frac{1}{2}}\right) d \leq \left(c^T \Sigma_{xx}^{-\frac{1}{2}} \Sigma_{xy} \Sigma_{yy}^{-\frac{1}{2}} \Sigma_{yy}^{-\frac{1}{2}} \Sigma_{yx} \Sigma_{xx}^{-\frac{1}{2}} c\right)^{\frac{1}{2}} \sqrt{d^T d},$$

where equality holds iff $X_{yy}^{-\frac{1}{2}} \Sigma yx \Sigma_{xx}^{-\frac{1}{2}} c$ and d are in the same direction. Hence in that case

$$\rho_1^2 = \frac{c^T \Sigma_{xx}^{-\frac{1}{2}} \Sigma_{xy} \Sigma_{yy}^{-\frac{1}{2}} \Sigma_{yy}^{-\frac{1}{2}} \Sigma_{yx} \Sigma_{xx}^{-\frac{1}{2}} c}{c^T c}. \tag{5}$$

Equation (5) is Rayleigh quatient, hence its maximizer is $c = v_1(\Sigma_{xx}^{-\frac{1}{2}} \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} \Sigma_{xx}^{-\frac{1}{2}})$. d is then obtained as $d = X_{yy}^{-\frac{1}{2}} \Sigma_{yx} \Sigma_{xx}^{-\frac{1}{2}} c$. Similarly, reversing the order of x and y in the above process we get that $d = v_1(\Sigma_{yy}^{-\frac{1}{2}} \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{yy}^{-\frac{1}{2}})$, and c is then obtained as $c = X_{xx}^{-\frac{1}{2}} \Sigma_{xy} \Sigma_{yy}^{-\frac{1}{2}} d$. Finally, reversing the change of variables we have $a = \Sigma xx^{-\frac{1}{2}} c$ and $b = \Sigma yy^{-\frac{1}{2}} d$. The projected variables are then $u_1 = a^T x$ and $v_1 = b^T y$. To obtain the next pairs $u_i, v_i, i = 2, \dots, \min m, n$, we would like each new canonical directions to be uncorrelated with previous ones, hence the subsequent eigenvectors are used (see Q2@homework).

Homework

1. Prove that $\max_{x:x^T v_1(A)=0} R(A,x) = \lambda_2(A)$.
2. Prove that the maximizer of the generalized Rayleigh quotient $R(A,x,y) = \frac{x^T A y}{\|x\|\|y\|}$ is $\sigma_1(A)$
3. PCA: Let X^T be a $n \times m$ data matrix. Express the projected data $X^T V$ in terms of the singular value decomposition of X^T
4. PCA: Show how to project the data onto $k < m$ first proncipal directions, and map it back to the original coordinates.
5. Let v, λ be an eigenpair of the the covariance $\mathbb{E}\left[xx^T\right]$. What is the variance along $x^T v$?
6. CCA: Compute the c,d using SVD (hint: use the previous question and equation (4)).
7. Programming: Create a collection of 10,000 natural images, resized to 32×32.
   (a) Compute PCA and show the leading 20 principal directions
   (b) Show reconstructions from $k = 5, 10, 20, 50, 100$ dimensions.
8. Programming:
   (a) Create point cloud in 2D on same manifold
   (b) Create a noisy copy of the cloud, by a linear transformation plus noise
   (c) use the fact that CCA returns matrices A, B such that $A^T x \approx B^T y$ to align the two point clouds. (Ok to use built in packages for CCA computation).