STA 5106
Computational Methods in Statistics I

Department of Statistics
Florida State University

Class 7
September 17, 2019

Review: Singular Value Decomposition

- **Theorem 2** For any $X \in \mathbf{R}^{m \times n}$ (assuming $m \geq n$) there exist orthogonal matrices $U \in \mathbf{R}^{m \times m}$ and $V \in \mathbf{R}^{n \times n}$ such that

$$U^T X V = \Sigma, \text{ where } \Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n) \in \mathbf{R}^{m \times n},$$

and where $\sigma_1 \geq \sigma_2 \geq \ldots, \sigma_n \geq 0$.

σ_i's are called the singular values of X and the columns of U and V are called the singular vectors of X.

$$X(v_1, v_2, \cdots, v_n) = (u_1, u_2, \cdots, u_m) \begin{pmatrix} \sigma_1 & 0 & & 0 \\ 0 & \sigma_2 & & 0 \\ & & \ddots & \\ 0 & 0 & & \sigma_n \\ 0 & 0 & 0 & 0 \\ & \cdots & \cdots & \end{pmatrix}$$

Review: Dimension Reduction

- If X happens to be a symmetric, semi-positive definite matrix then the orthogonal matrices U and V are the same.
- If the original observation space is $\mathbf{R}^D$, then the problem reduces to finding an appropriate projection that takes elements of $\mathbf{R}^M$ to elements of $\mathbf{R}^D$ ($M < D$) in a linear fashion.
- The **principal components** of x are given by $z = U^T x$ where $U \Lambda U^T$ is the SVD of the covariance of x.
- Let the elements of Λ be non-increasing from top-left to bottom-right. i.e. $\boldsymbol{\lambda}_1 \geq \boldsymbol{\lambda}_2 \geq \ldots \geq \boldsymbol{\lambda}_D$. In general, z_i is uncorrelated to the other principal components, and has the variance $\boldsymbol{\lambda}_i$.

Review: PCA Algorithm

- **Algorithm 20 (PCA of Given Data)** Let X be the $D \times n$ matrix where **each column** denotes an independent observation vector for the random vector x.

  1. Find the sample covariance matrix $C \in \mathbf{R}^{D \times D}$ of the elements of X,

  2. Compute the singular value decomposition (SVD) of C to obtain the orthogonal matrix $U \in \mathbf{R}^{D \times D}$,

  3. Set U_1 to be the first M columns of U, and,

  4. define $Z = U_1^T X \in \mathbf{R}^{M \times n}$.

Review: One-Dimensional Projection

- Given $x_1, \ldots, x_N$ in $\mathbf{R}^D$, let the projection vector be u_1, then the projected value is

$$u_1^T x_n, \qquad n = 1, \ldots, N$$

- Using Lagrange multiplier, we can maximize

$$u_1^T S u_1 + \lambda(1 - u_1^T u_1)$$

- By taking derivative with respect to u_1, we have

$$S u_1 = \lambda u_1.$$

- Moreover, we see that the variance is given by $u_1^T S u_1 = \lambda.$ So the variance will be a maximum when we set $\boldsymbol{u_1}$ equal to the eigenvector having the **largest eigenvalue**.

Matrix Calculus

- Definition: assume $X = (x_{ij})_{mn}$, then

$$d/dX = \begin{pmatrix} d/dx_{11} & \cdots & d/dx_{1n} \\ \vdots & \ddots & \vdots \\ d/dx_{m1} & \cdots & d/dx_{mn} \end{pmatrix}$$

- Quadratic Products:

$$d/d\mathbf{X}\ ((\mathbf{Xa+b})^T\mathbf{C}(\mathbf{Xa+b})) = (\mathbf{C}+\mathbf{C}^T)(\mathbf{Xa+b})\mathbf{a}^T$$

(upper case: matrix, lower case: column vector)

- Therefore,

$$\frac{\partial[u_1^T S u_1 + \lambda(1 - u_1^T u_1)]}{\partial u_1} = 2Su_1 - 2\lambda u_1$$

http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/calculus.html

Review: Multi-Dimensional Projection

- **The optimal linear projection for which the variance of the projected data is maximized is now defined by the M eigenvectors $u_1, \ldots, u_M$ of the data covariance matrix S corresponding to the M largest eigenvalues $\lambda_1, \ldots, \lambda_M$.**

- Let the projection matrix be $V = [v_1,\ldots,v_M,\ldots, v_D]$. This is a orthogonal matrix in $\mathbf{R}^{D\mathrm{x}D}$, i.e. $VV^T = V^TV = I_D$.

- The submatrix $B = [v_1,\ldots,v_M]$ with $B^TB = I_M$. Then the projected value is

$$B^T x_n, \qquad n = 1, \ldots, N$$

- We have shown that

$$B^T x_n = C[u_1, \ldots, u_M]^T x_n = C[u_1^T x_n, \ldots, u_M^T x_n]^T.$$

Linear Discriminant Analysis

- In case the goal is to separate and categorize the observations after their projection, then the choice of projection takes into account this categorization.

- In other words, we can try to choose a projection that minimizes the spread of projected vectors within the same class, and maximizes the separation of vectors across different classes.

- In this section we derive an optimal projection using such a criterion.

Optimal Projection

- Let $C_1, C_2, \ldots, C_m$ be m sets that partition $\mathbf{R}^n$ into m classes.

- We are given N_j observations for class labelled by C_j, i.e. given

$$X_i^j \in C_j,\ i = 1, \ldots, N_j, \text{ for } j = 1, \ldots, m.$$

where each X_i^j is an element in $\mathbf{R}^n$.

- The goal is to find a projection, denoted by the basis U, such that the coefficients from the same class cluster closer and the classes separate away maximally.

Useful Quantities

- Let μ_j be the mean of observations in class C_j:

$$\mu_j = \frac{1}{N_j}\sum_{i=1}^{N_j} X_i^j \in \mathbf{R}^n$$

- A matrix that captures the separation between the classes is the between-class scatter matrix:

$$S_B = \sum_{j=1}^{m}(\mu_j - \mu)(\mu_j - \mu)^T \in \mathbf{R}^{n\times n}, \qquad \text{where } \mu = \frac{1}{m}\sum_{j=1}^{m}\mu_j$$

- Similarly, the separation between elements within the same class is captured by the within-class scatter matrix:

$$S_W = \sum_{j=1}^{m}\left(\sum_{i=1}^{N_j}(X_i^j - \mu_j)(X_i^j - \mu_j)^T\right) \in \mathbf{R}^{n\times n}$$

Maximization Problem

- In case $Z = U^T X \in \mathbf{R}^k$ is the projection of X onto the space spanned by the columns of U ($\in \mathbf{R}^{n\times k}$), then the scatter matrices for the projections are derived similarly.

- The corresponding scatter matrices are given by:

$$S_B^z = U^T S_B U, \text{ and } S_W^z = U^T S_W U$$

- The goal is to choose U that maximizes the following functions:

$$f(U) = \frac{\det(S_B^z)}{\det(S_W^z)} = \frac{\det(U^T S_B U)}{\det(U^T S_W U)}$$

Optimal Solution

- The optimal projection is given by:

$$\hat{U} = \underset{U \in \mathbf{R}^{n \times k}, U^T U = I_k}{\arg\max} f(U)$$

- It can be shown that the solution can be solved as the generalized eigen-value problem.

- In other words, the columns of $\hat{U}$ satisfy the equation:

$$S_B \hat{u}_i = \lambda_i S_W \hat{u}_i.$$

- In Matlab, this can be achieved using the **eig** function for solving the generalized eigen-value problem.