LECTURE 6 - PRINCIPAL COMPONENTS ANALYSIS

1. Lagrangian Optimization - Equality Constraints

- Lets consider constrained optimization with *equality constraints*
- Lagrange multipliers are auxiliary variables that help to characterize optimal solutions. They can be viewed in two ways:
  - **Penalty Viewpoint:** In this method, we disregard the constraints but add a very high penalty for violating them. Then, we can work with the "penalized" unconstrained problem
  - **Feasible Direction Viewpoint:** This method relies on the fact that at a local minimum, there can be no cost improvement when traveling a small distance angle a direction that leads to feasible solutions
- Consider problems with equality constraints of the following form:

$$\min_x f(x) \text{ such that } h_i(x) = 0, \quad i = 1, \ldots, m \tag{1}$$

We assume that f and h_i are continuously differentiable functions.

- The basic Lagrange multiplier theorem states that for a given local minimum, x^*, there exist scalars $\lambda_1, \ldots, \lambda_m$ called *Lagrange Multipliers* such that,

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) = 0 \tag{2}$$

There are two ways to interpret this equation:

- The cost gradient $\nabla f(x^*)$ belongs to the subspace spanned by the constraint gradients at x^*
- The cost gradient $\nabla f(x^*)$ is orthogonal to the subspace of the *first order feasible variations*

$$V(x^*) = \left\{ \Delta x | \nabla h_i(x^*)' \Delta x = 0, i = 1, \ldots, m \right\} \tag{3}$$

- For a geometric perspective, note that at any point on the constraint surface the gradient $\nabla g(\mathbf{x})$ of the constraint function will be orthogonal to the surface. Consider a point $\mathbf{x}$ that is on the constraint surface and consider a nearby point $\mathbf{x}+\epsilon$ also on the surface. Then the Taylor series expansion around $\mathbf{x}$ is: $g(\mathbf{x}+\epsilon) \approx g(\mathbf{x}) + \epsilon^T \nabla g(\mathbf{x})$. Since both $\mathbf{x}$ and $\mathbf{x}+\epsilon$ are on the constraint surface, then $g(\mathbf{x}+\epsilon) = g(\mathbf{x})$ and $\epsilon^T \nabla g(\mathbf{x}) \approx 0$. As $\|\epsilon\| \to 0$, then $\epsilon^T \nabla g(\mathbf{x}) = 0$ because ϵ is parallel to the constraint surface and $\nabla g(\mathbf{x})$ is orthogonal to the surface.
- Also, for any point in which $f(\mathbf{x})$ is maximized along the constraint surface, then $\nabla f(\mathbf{x})$ is orthogonal to the constraint surface. Otherwise, then there would be a step along the constraint surface in which $f(\mathbf{x})$ could be further increased.

- Hence, $\nabla f(x^*) + \lambda \nabla g(x^*) = 0$
- For additional reading on Lagrangian optimization, see: Appendix E in the Bishop textbook and/or Constrained Optimization and Lagrange Multiplier Methods by Bertsekas, `http://www.mit.edu/~dimitrib/Constrained-Opt.pdf`

1.1. **Principal Components Analysis - Maximal Variance Formulation.**

- PCA is a linear transformation
- PCA minimizes the redundancy of the resulting transformed data (by ending up data that is uncorrelated), minimizes the mean squared error between original and transformed/reduced data, and maximizes the retained variance of the data.
- Consider a data set of observations $\{\mathbf{x}_n\}_{n=1}^N$ and $\mathbf{x}_n \in \mathbb{R}^D$. We want to maximize the variance of the projected data.
- Let us first consider reducing dimensionality to $M = 1$. Let us define the projection as a vector $\mathbf{u}_1$ where $\mathbf{u}_1^T\mathbf{u}_1 = 1$. Then, each projected data point into 1-D would be $y_n = \mathbf{u}_1^T\mathbf{x}_n$
- The mean of the sample data is $\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^N \mathbf{x}_n$ and the mean of the projected data is $\mathbf{u}_1^T\bar{\mathbf{x}}$
- The variance of projected data is:

$$\frac{1}{N}\sum_{n=1}^{N}\left\{\mathbf{u}_1^T\mathbf{x}_n - \mathbf{u}_1^T\bar{\mathbf{x}}\right\}^2 \quad (4)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{u}_1^T\mathbf{x}_n - \mathbf{u}_1^T\bar{\mathbf{x}}\right)\left(\mathbf{u}_1^T\mathbf{x}_n - \mathbf{u}_1^T\bar{\mathbf{x}}\right)^T \quad (5)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{u}_1^T\mathbf{x}_n - \mathbf{u}_1^T\bar{\mathbf{x}}\right)\left(\mathbf{x}_n^T\mathbf{u}_1 - \bar{\mathbf{x}}^T\mathbf{u}_1\right) \quad (6)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mathbf{u}_1^T\mathbf{x}_n\mathbf{x}_n^T\mathbf{u}_1 - \mathbf{u}_1^T\mathbf{x}_n\bar{\mathbf{x}}^T\mathbf{u}_1 - \mathbf{u}_1^T\bar{\mathbf{x}}\mathbf{x}_n^T\mathbf{u}_1 + \mathbf{u}_1^T\bar{\mathbf{x}}\bar{\mathbf{x}}^T\mathbf{u}_1 \quad (7)$$

$$= \mathbf{u}_1^T\left(\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T - \mathbf{x}_n\bar{\mathbf{x}}^T - \bar{\mathbf{x}}\mathbf{x}_n^T + \bar{\mathbf{x}}\bar{\mathbf{x}}^T\right)\mathbf{u}_1 \quad (8)$$

$$= \mathbf{u}_1^T\left(\frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n - \bar{\mathbf{x}}\right)\left(\mathbf{x}_n - \bar{\mathbf{x}}\right)^T\right)\mathbf{u}_1 \quad (9)$$

$$= \mathbf{u}_1^T\mathbf{S}\mathbf{u_1}$$

- Now, we can maximize the projected variance with respect to $\mathbf{u}_1$ while constraining $\mathbf{u}_1^T\mathbf{u}_1 = 1$. We will do this using a Lagrange multiplier:

$$L = \mathbf{u}_1^T\mathbf{S}\mathbf{u_1} + \lambda_1\left(1 - \mathbf{u}_1^T\mathbf{u}\right) \quad (10)$$

- By taking the derivative of the Lagrangian and setting it equal to zero, we get:

$$\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u_1} \quad (11)$$

- We can left multiply by $\mathbf{u}_1^T$ and get:

$$\mathbf{u_1}^T\mathbf{S}\mathbf{u_1} = \lambda_1 \tag{12}$$

- So the variance of the projected data is equal to the eigenvalue of the covariance matrix of the sample data along the direction of the eigenvector used for dimensionality reduction.
- We can incrementally add new eigenvector directions (ordered by maximal eigenvalue/variance) to project into an M dimensional space where $1 \leq M \leq D$

1.2. **PCA for Minimization of Mean Squared Error.**

- We can also look at PCA as a minimization of mean squared error.
- Consider $\mathbf{x} \in R^n$ and an orthogonal basis $\mathbf{a}$:

$$\hat{\mathbf{x}} = \sum_{i=1}^{m} y_i \mathbf{a}_i \tag{13}$$

where $m < n$.

$$y_j = \mathbf{x}^T \mathbf{a}_j \tag{14}$$

where $\mathbf{A}^T\mathbf{A} = \mathbf{I}$

We want to minimize the residual error:

$$\epsilon = \mathbf{x} - \hat{\mathbf{x}} = \sum_{j=m+1}^{n} y_j \mathbf{a}_j \tag{15}$$

The objective we will used is the mean square residual:

$$J = E\{\|\epsilon\|_2^2\} \tag{16}$$

$$= E\left\{\left(\sum_{i=m+1}^{n} y_i \mathbf{a}_i^T\right)\left(\sum_{i=m+1}^{n} y_i \mathbf{a}_i\right)\right\} \tag{17}$$

$$= \sum_{j=m+1}^{n} E\{y_j^2\} \tag{18}$$

$$= \sum_{j=m+1}^{n} E\{(\mathbf{a}_j^T\mathbf{x})(\mathbf{x}^T\mathbf{a}_j)\} \tag{19}$$

$$= \sum_{j=m+1}^{n} \mathbf{a}_j^T E\{\mathbf{x}\mathbf{x}^T\}\mathbf{a}_j \tag{20}$$

$$= \sum_{j=m+1}^{n} \mathbf{a}_j^T R_x \mathbf{a}_j \tag{21}$$

Minimize the error and incorporate Lagrange parameters for $\mathbf{A}^T\mathbf{A} = \mathbf{I}$:

$$\frac{\partial J}{\partial \mathbf{a}_j} = 2(R_x\mathbf{a}_j - \lambda_j\mathbf{a}_j) = 0 \text{ for } j = m+1 \ldots n \tag{22}$$

$$R_x\mathbf{a}_j = \lambda_j\mathbf{a}_j \tag{23}$$

So, the sum of the error is the sum of the eigenvalues of the unused eigenvectors. So, we want to select the eigenvectors with the m largest values.