Linear Algebra Working Group :: Day 3 Part 1

Note: All vector spaces will be finite-dimensional vector spaces over the field $\mathbb{R}$.

1 Principal component analysis and dimensional reduction

Definition 1.1. Given an $m \times N$ matrix X of observations, with columns X_j thought of as m-dimensional observation vectors, the **sample mean** of the observation vectors is the vector:

$$M := \frac{1}{N} \sum_{j=1}^{N} X_j$$

The matrix:

$$B := \begin{pmatrix} X_1 - M & X_2 - M & \dots & X_N - M \end{pmatrix}$$

is called the **mean-deviation form** of the matrix X of observations. The columns of the matrix B are often denoted by $\widehat{X}_j$.

Remark 1.2. When we think of a matrix of observations X, one can think of the columns X_j as one set of observations of m variables. Thus, the rows correspond to variables and the row X^i of the matrix are N observations of the i^{th} variable. In these applications, N is usually large.

Exercise 1. Given an $m \times N$ matrix X of observations, show that its mean deviation form has zero sample mean.

Definition 1.3. Given a vector of observations $x = (x_1, ..., x_N)$, let $\widehat{x}$ be the average of the observations. The **sample variance** of the observations is the quantity:

$$\text{Var}(x) := \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \widehat{x})^2$$

Definition 1.4. Given two vectors of observations $x = (x_1, ..., x_N)$ and $y = (y_1, ..., y_K)$ with means $\widehat{x}$ and $\widehat{y}$ respectively. The (sample) **covariance** of the observations is the quantity:

$$\text{Covar}(x, y) := \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \widehat{x})(y_i - \widehat{y})$$

When the covariance between x and y is 0 we say the data x and y are **uncorrelated**.

Definition 1.5. Given an $m \times N$ matrix of observations X, let B be the mean-deviation form of X. The **sample-covariance matrix** of the matrix of observations is the $m \times m$matrix S defined by:

$$S := \frac{1}{N-1} BB^T$$

Exercise 2. Show that the sample covariance matrix of a matrix of observations is positive semidefinite (Use exercise 24 or 31 from Day 2.)

Exercise 3. Let X be a matrix of observations and let S be the covariance matrix of X. Suppose that the matrix X is already in mean-deivation form. Show that the diagonal entry S_{ii} of the matrix S corresponds to the variance of the i^{th} row of X viewed as a vector of observations. Show that the off-diagonal entry S_{ij} of the covariance matrix corresponds to the covariance of the i^{th} and j^{th} row of X.

Definition 1.6. Let X be an $m \times N$ matrix of observations and S its covariance matrix. The **total variance** of the observation matrix X is the trace:

$$\text{tr}(S) := \sum_{j=1}^{m} S_{jj}$$

Exercise 4. Let A and B be two $n \times n$ matrices.

1. Show that $\text{tr}(AB) = \text{tr}(BA)$.
2. Show that if A and B are similar, $\text{tr}(A) = \text{tr}(B)$.

Exercise 5. Let X be an $m \times N$ matrix of observations already in mean-deviation form and let P be an $m \times m$ orthogonal matrix. Let Y be the $m \times N$ matrix $Y := P^T X$. Then:

1. Show that the matrix Y is in mean-devation form.
2. Show that if the covariance matrix of X is S, then the covariance matrix of Y is $P^T S P$.
3. Show that the total variances of X and Y are the same.

Definition 1.7. Let X be an $m \times N$ matrix of observations and let S be its covariance matrix. Let S have the eigenvalues:

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m \geq 0$$

with corresponding orthonormal eigenvectors $u_1, u_2, ..., u_m$. The eigenvector u_i is called the $i^{\textbf{th}}$ **principal component** of the data.

Remark 1.8. Principal component analysis consists of taking a matrix of observations X and finding an orthogonal change of variables $Y = P^T X$ that makes the new variables uncorrelated. The reason for requiring orthogonality can be seen in 5. We also want to put them in order of decreasing variance for the sake of choosing a convention.

Exercise 6. Let X be a matrix of observations in mean-deviation form and let S be its covariance matrix. Use facts about symmetric matrices from Day 2, to show there exists an orthogonal change of variables $Y = P^T X$ such that the matrix P consists of the principal components and the new covariance matrix shows the new variables are uncorrelated.

Exercise 7. Consider the following matrix of observations:

$$X = \begin{pmatrix} 19 & 22 & 6 & 3 & 2 & 20 \\ 12 & 6 & 9 & 15 & 13 & 5 \end{pmatrix}$$

1. Convert the matrix of observations to mean-deviation form.
2. Construct the sample covariance matrix.

3. Find the principal components of the data.

4. Perform a change of variables to principal components.

Remark 1.9. Given a matrix of observations, dimensional reduction consists of performing a change of variables to principal components and then orthogonally projecting to the subspace with the "overlwhelming" amount of variance.

Exercise 8. Suppose a 3×1000 matrix of observations X has the following covariance matrix:

$$S = \begin{pmatrix} 70 & 0 & 0 \\ 0 & 20 & 5\sqrt{3} \\ 0 & 5\sqrt{3} & 10 \end{pmatrix}$$

1. Obtain the principal components.

2. In the new variables, what are the proportions of each of the variances to the total variance?

3. Should we do dimensional reduction? To which subspace should we project?

4. Starting from the matrix of observations X, what does it mean to reduce dimensions as in remark 1.9? That is, what observations do we consider when we perform dimensional reduction on X?

Exercise 9. Let X be an $m \times N$ matrix of observations. Let $A := \frac{1}{\sqrt{N-1}} X^T$. Suppose $A = U\Sigma V^T$ is a singular value decomposition of A. Identify the eigenvalues of the covariance matrix and the principal components from the singular value decomposition.

References

[Hal58] P.R. Halmos. Finite-Dimensional Vector Sapces. Reprinting of the 1958 second edition. *Undergraduate Texts in Mathematics*. Springer-Verlag, New York-Heidelberg, 1974.

[L94] D.C. Lay. Linear Algebra and its Applications. Fourth Edition. Addison-Wesley, 2012.

[Rom08] S. Roman. Advanced Linear Algebra. Third Edition. *Graduate Texts in Mathematics, 135*. Springer, New York, 2008.