CISC 271 Class 16

Principal Components Analysis – PCA

Text Correspondence: 7.3

Main Concepts:

- *Zero-mean data: average value is zero*
- *Covariances of differences: symmetric positive definite matrix*
- *Principal components: Eigenvectors of the covariances*
- *Reconstructed data: data mean plus selected principal components*

Sample Problem, Data Analysis: For a set of data vectors, what vector space captures "most" of the variance of the data?

16.1 Motivation, by Example

Empirical data in matrix form, such as in a table or spreadsheet, typically have distinct meanings for rows and columns. For example, an instructor in a fictitious course might have tabulated grades for tests. Each column would be the grades of an individual student, and each row would be the grades achieved in a particular test. The data might look like

	Ami	Ben	Cindy	Davis	$\cdots$	Zoe
Test 1	30	20	31	28	$\cdots$	33
Test 2	22	18	22	23	$\cdots$	30
Test 3	28	19	27	28	$\cdots$	36
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

An instructor might want to know how well the students are doing on average, and how much students vary from the average. These questions can be address by finding the main, or *principal*, ways that each column of data vary from the average column.

The average performance is simply the average for each test. The ways that students vary from each test are harder to assess. In this example, most students seem to do better on Test #1 than on Test #2, and Test #3 seems to be somewhere between the first two tests. We want to do data analysis that captures systematic variations, if this is possible.

The data in the above table are not in the form that we are using in this course. We require that each column contains the values of a single variable, and that each row contains the values

of observations. We would need to transpose the above table into a data matrix. For the above example, the data matrix might be described by a matrix A_1 as

$$A_1 = \begin{bmatrix} 30 & 22 & 28 \\ 20 & 18 & 19 \\ 31 & 22 & 27 \\ 28 & 23 & 28 \\ \vdots & \vdots & \vdots \\ 33 & 30 & 36 \end{bmatrix} \tag{16.1}$$

The problem of finding the principal ways that the variables differ from the mean is called *principal components analysis*, or PCA.

16.2 Zero-Mean Data Matrix

Referring to the motivating example, the average student performance can be computing by finding the average mark on each test. For the data matrix A_1 of Equation 16.1, this can be done by performing the first step in data standardization: subtract, from each column, the mean value of that column. We can write this *zero-mean* data matrix as M, so for the example in Equation 16.1 the zero-mean data matrix would be

$$M_1 = A_1 - \vec{1}\,\bar{A} = \begin{bmatrix} \vec{m}_1 & \vec{m}_2 & \cdots & \vec{m}_n \end{bmatrix} = \begin{bmatrix} 1.60 & -1.00 & 0.40 \\ -8.40 & -5.00 & -8.60 \\ 2.60 & -1.00 & -0.60 \\ -0.40 & 0.00 & 0.40 \\ 4.60 & 7.00 & 8.40 \end{bmatrix} \tag{16.2}$$

We will note here that this not a universal convention for writing the zero-mean matrix. For example, the textbook uses a column to represent an observation, so the textbook would write the zero-mean matrix as

$$[M_1]^T$$

We are using the `MATLAB` convention of Equation 16.2 so that the class notes correspond more closely with our code. This is a reminder that the representation convention must be carefully understood before a resource is consulted.

16.3 Principal Components Analysis as an SVD

In statistics, the principal components are derived from the sample covariance matrix of the zero-mean data. Using our notation, this would be a symmetric positive semidefinite matrix B. Here, again, we must be careful to observe that we are using *sample* statistics and not population statistics. In this course, the covariance matrix is defined as

$$B = \left(\frac{1}{m-1}\right) M^T M \tag{16.3}$$

For our example data, the covariance matrix would be

$$B_1 = \frac{M_1^T M_1}{m-1} = \begin{bmatrix} 25.30 & 17.50 & 27.45 \\ 17.50 & 19.00 & 25.50 \\ 27.45 & 25.50 & 36.30 \end{bmatrix} \tag{16.4}$$

The principal components of a data matrix M are the eigenvectors of its covariance matrix B. To two digits of numerical precision, the eigenvalue vector $\vec{\lambda}$ and the eigenvector matrix E_1 are

$$B_1 = E_1 \Lambda_1 E_1^T \quad \text{where} \quad \vec{\lambda} = \begin{bmatrix} 75.60 \\ 4.82 \\ 0.17 \end{bmatrix} \qquad E_1 = \begin{bmatrix} 0.54 & 0.79 & 0.29 \\ 0.48 & -0.58 & 0.66 \\ 0.69 & -0.22 & -0.69 \end{bmatrix} \tag{16.5}$$

From a statistics point of view, Equation 16.5 tells us that "most" of the variance in the student grades are captured by a single component. A minor amount of variance is captured by using a second component and the third component can be numerically neglected: it constitutes only 0.22% of the overall variance.

Now, let us examine the covariance matrix by using the SVD. We know, from previous classes, that the SVD is closely related to the spectral decomposition of the covariance matrix as we have defined it. The SVD of any zero-mean data matrix M is

$$M = U \Sigma V^T \tag{16.6}$$

For a data matrix with m rows, the spectral decomposition of the covariance matrix is

$$B = \frac{M^T M}{m-1} = \frac{E \Lambda E^T}{m-1} = E \frac{\Lambda}{m-1} E^T \tag{16.7}$$

A relationship between the SVD of a zero-mean matrix and the spectral decomposition of its covariance matrix can be deduced by expanding Equation 16.6 into Equation 16.7, which is

$$B = \frac{M^T M}{m-1} = \frac{V \Sigma V^T V \Sigma V^T}{m-1} = V \frac{\Sigma^2}{m-1} V^T = E \frac{\Lambda}{m-1} E^T \tag{16.8}$$

From Equation 16.8, we can verify our previous finding that the right singular matrix of the SVD of a zero-mean data matrix M is the same as the eigenvector matrix of the covariance matrix: $E = V$. We can also verify the relationship between the singular values of a zero-mean data matrix and the eigenvalues of its covariance matrix: $\lambda_j = \sigma_j^2/(m-1)$.

16.4 Using the SVD to Compute PCA Scores

Equation 16.8 informs us that we can find principal components of a data matrix by using the SVD. There is some artistry, or human intelligence, that may be involved in selecting the number of relevant components. For our example of grades in a class, we might explore using one or two components of the data.

We will define the word *component* as a unit-length eigenvector of the covariance matrix B. Because of the definition of the SVD, we can equally well define – or simply use – a right singular vector of the zero-mean data matrix M.

After the number of components are chosen, there are two common uses of PCA:

1. Score the PCA to reduce the dimensionality of the data
2. Reconstruct the data from the PCA

In this course we will explore the first use, recognizing that the second use also has many applications.

The idea of a PCA *score* of data is to project a zero-mean observation onto a unit-length principal component. The first score of the i^{th} observation is the product of the i^{th} row of the zero-mean matrix M and the first right singular vector $\vec{v}_1$. If we write the i^{th} row of M as $\vec{t}_i^{\,T}$, then our computation for the i^{th} observation in the original data matrix A will find the j^{th} score s_{ij} that corresponds to the j^{th} principal component $\vec{v}_j$ as

$$z_{ij} = \vec{z}_i^{\,T} \vec{v}_j \tag{16.9}$$

We can perform all of the scoring in Equation 16.9 – that is, find every individual score z_{ij} – with the single matrix multiplication

$$Z = MV \tag{16.10}$$

The matrix Z in Equation 16.10 will contain the scores of the principal components of the data matrix A.

For our example of student grades in a class, the zero-mean data are provided numerically in Equation 16.2 and the principal components are provided in Equation 16.5, as the matrix $E_1 = V_1$. If we use the first two components, then the score matrix Z_1 can be found as

$$Z_1 = \begin{bmatrix} \vec{z}_1 & \vec{z}_2 \end{bmatrix} = M_1 \begin{bmatrix} \vec{v}_1 & \vec{v}_2 \end{bmatrix} = \begin{bmatrix} -0.67 & 1.75 \\ 12.89 & -1.84 \\ -0.52 & 2.76 \\ -0.06 & -0.40 \\ -11.64 & -2.26 \end{bmatrix} \tag{16.11}$$

We can see numerical patterns in the entries of the score vector $\vec{z}_1$, with three entries being close to zero and two entries being quite far from zero. We cannot deduce anything from the $\pm$ sign of the entries because the signs of eigenvector entries can be sensitive to numerical details.

The entries of the score vector $\vec{z}_2$ do not have a distinctive pattern. The entries of $\vec{z}_2$ are, mostly, about one standard deviation from zero.

These patterns are consistent with the observation about the magnitudes of the eigenvalues of the covariance matrix B_1 for this example. Most of the variance was captured by the first principal component and much less was captured by the second principal component.

We can visualize these results by using a "scatter" plot. The first axis, horizontal, is the score of each observation for the first principal component. The second axis, vertical, is the score of each observation for the second principal component. Figure 16.1 illustrates the PCA results for the example data matrix A_1.

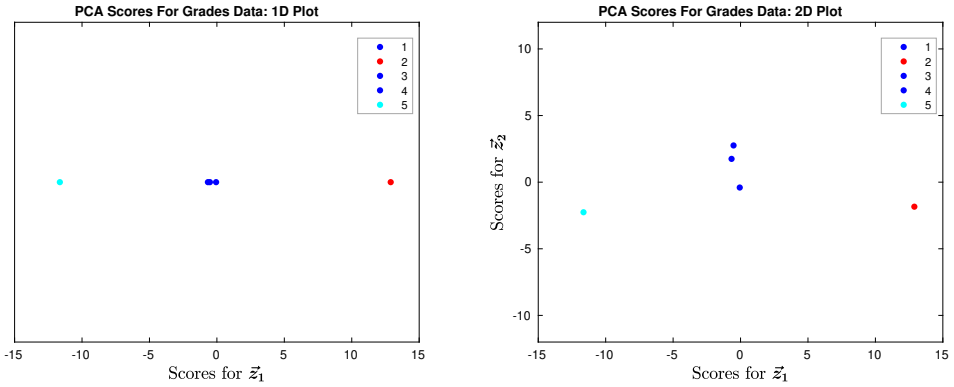

Figure 16.1: PCA scatter plot for the example data of grades in the matrix A_1. Data are colored to indicate potential clusters of the data. (A) The horizontal axis is the score for the first principal component. (B) The vertical axis is the score for the second principal component.

We can visually cluster the data in Figure 16.1, which are rows in the score matrix Z_1. It is plausible to assign three of the observations to a central cluster, and each of the other two observations to distinct cluster. Visually, it is unclear that using the second component provides us with information about the data. For both numerical reasons and visual reasons, we might choose to represent the data matrix A_1 by a single principal component.

This example shows how PCA can be used to perform *dimensionality reduction*. Here, we reduced the number of dimensions of the data from the original number of variables – which was 3 – to the final number of principal components, which was 1. This choice communicates to our readers how much information is needed to cluster the observation in the data matrix A_1. In this small example, we can identify three of the students as having the same performance; one students significantly under-performed on the tests and one student out-performed on the tests.

Extra Notes

One of the many uses of the SVD is that its formulation produces a *series* for any given matrix [3]. To understand how we can use such a series in linear data analysis, we will first explore the norm of a matrix.

16.5 Matrix Norms

A *norm* of a vector or matrix is a real number that "measures" the object. We write the vector norm as $\|\vec{a}\|$ and the matrix norm as $\|A\|$; a common abbreviation for the norm of an object is $\|\cdot\|$. A norm must satisfy four axioms, which we will write for a matrix. The first four axioms are the same as the axioms for a vector norm and some authors add a fifth axiom for matrix norms: the object being measured has changed.

For any $A \in \mathbb{R}^{m\times n}$, any $A \in \mathbb{R}^{m\times n}$, and any $\alpha \in \mathbb{R}$, the axioms of a norm are:

- $\|A\| \geq 0$
- $\|A\| = 0$ if and only if $A = 0$
- $\|\alpha A\| = |\alpha|\ \|A\|$
- $\|A + C\| \leq \|A\| + \|C\|$
- $\|AC\| \leq \|A\|\ \|C\|$ (not universally required)

Sometimes, we want a vector norm and a matrix norm to "work" together.

Definition: compatible norms

For any vector norm $\|\cdot\|$,any matrix norm $\|\cdot\|$, any vector $\vec{w} \in \mathbb{R}^n$, and any matrix $A \in \mathbb{R}^{m \times n}$, the norms are *compatible* is defined as

$$\|A\vec{w}\| \leq \|A\| \, \|\vec{w}\| \tag{16.12}$$

The vector norm that we use in this course is the Euclidean norm. It is also written as the ℓ^2 norm, pronounced *ell–2*, which is the abbreviation that we will prefer. For a vector $\vec{w} \in \mathbb{R}^n$, this norm is defined as

$$\|\vec{w}\|_2 \stackrel{\text{def}}{=} \sqrt{\sum_{j=1}^{n} (w_j)^2} \tag{16.13}$$

The ℓ^2 matrix norm is an *induced* norm, which means that it is based on the vector norm. For a matrix $A \in \mathbb{R}^{m \times n}$, the ℓ^2 norm is

$$\|A\|_2 \stackrel{\text{def}}{=} \max_{\vec{w} \neq \vec{0}} \frac{\|A\vec{w}\|}{\|\vec{w}\|} \tag{16.14}$$

The *Frobenius* norm is another matrix norm that is closely related to the Euclidean vectorm norm. It is defined as

$$\|A\|_F \stackrel{\text{def}}{=} \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} (a_{ij})^2} \tag{16.15}$$

16.6 Eigenvalues and Singular Values

The above norms are closely related to matrix decompositions that we explored earlier in the course. These relations are theorems in linear algebra that can be found in many textbooks and other sources.

Abbreviate the largest eigenvalue of a matrix as $\lambda_{MAX}(\cdot)$ and recall the SVD of a matrix, in which the singular values are ordered as $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_r > 0$. The above matrix norms have the properties

$$\begin{aligned} \|A\|_2 &= \sqrt{\lambda_{MAX}(A^T A)} \\ &= \sigma_1 \\ \|A\|_F &= \sqrt{\sum_{j=1}^{r} \sigma_j^2} \end{aligned} \tag{16.16}$$

16.7 Extra Notes: A Matrix As A Series

Let us recall the SVD of a matrix $A \in \mathbb{R}^{m \times n}$ that is rank r. Because we can neglect each row and each column that has an index greater that r, we can write this matrix as

$$\begin{aligned} A &= U\Sigma V^T \\ \text{where} \quad U &= \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_r \end{bmatrix} \\ \Sigma &= \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix} \\ V^T &= \begin{bmatrix} \vec{v}_1^T \\ \vec{v}_2^T \\ \vdots \\ \vec{v}_r^T \end{bmatrix} \end{aligned} \tag{16.17}$$

We can expand the right product of Equation 16.17 as

$$\begin{aligned} \Sigma V^T &= \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix} \begin{bmatrix} \vec{v}_1^T \\ \vec{v}_2^T \\ \vdots \\ \vec{v}_r^T \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1 \vec{v}_1^T \\ \sigma_2 \vec{v}_2^T \\ \vdots \\ \sigma_r \vec{v}_r^T \end{bmatrix} \end{aligned} \tag{16.18}$$

Using Equation 16.18, we can write the SVD of Equation 16.17 as

$$\begin{aligned} A &= U\Sigma V^T \\ &= \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_r \end{bmatrix} \begin{bmatrix} \sigma_1 \vec{v}_1^T \\ \sigma_2 \vec{v}_2^T \\ \vdots \\ \sigma_r \vec{v}_r^T \end{bmatrix} \\ &= \vec{u}_1 \sigma_1 \vec{v}_1^T + \vec{u}_2 \sigma_2 \vec{v}_2^T + \cdots + \vec{u}_r \sigma_r \vec{v}_r^T \\ &= \sigma_1 [\vec{u}_1 \vec{v}_1^T] + \sigma_2 [\vec{u}_2 \vec{v}_2^T] + \cdots + \sigma_r [\vec{u}_r \vec{v}_r^T] \end{aligned} \tag{16.19}$$

16.8 Extra Notes: Eckart-Young Theorem

Equation 16.19 is a remarkable series. It states that any rank–r matrix A can be represented as the sum of r rank–1 matrices. Each rank–1 matrix in the series is the product of the corresponding left singular vector and right singular vector of A.

The SVD was observed and discovered, at least partially, since the late 19^{th} century. It was formulated and shown to exist in 1936, by Eckart and Young [3]; their approximation is known in linear algebra as the Eckart-Young Theorem.

The approximation theorem has been stated and proved for many matrix norms. We can use either the ℓ^2 norm or the Frobenius norm. For a matrix A, we will define the i^{th} rank–1 matrix in Equation 16.19 as

$$C_i \stackrel{\text{def}}{=} \sigma_i[\vec{u}_i \vec{v}_i^T] \tag{16.20}$$

Formally, the Eckart-Young theorem states that the optimal rank–p approximation to the matrix A is the first p terms of the series in Equation 16.19. For $A \in \mathbb{R}^{m \times n}$, the optimal approximation is

$$\begin{aligned} C &= \underset{W \in \mathbb{R}^{m \times n}}{\text{argmin}} \, \|A - W\|_2 \\ &= \underset{W \in \mathbb{R}^{m \times n}}{\text{argmin}} \, \|A - W\|_F \\ &= C_1 + C_2 + \cdots + C_p \end{aligned} \tag{16.21}$$

Using Equation 16.21 and Equation 16.17, we can find the SVD of the optimal rank–p approximation to A as

$$\begin{aligned} C &= U_p \Sigma_p V_p^T \\ \text{where} \quad U_p &= \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_p \end{bmatrix} \\ \Sigma_p &= \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p \end{bmatrix} \\ V_p^T &= \begin{bmatrix} \vec{v}_1^T \\ \vec{v}_2^T \\ \vdots \\ \vec{v}_p^T \end{bmatrix} \end{aligned} \tag{16.22}$$

We use Equation 16.22 often when we perform linear data analysis. One way that we can think of this approximation is:

The optimal rank–p approximation to the column space of A is U_p

That is, when we want to use a smaller set of basis vectors to span the column space of a matrix A, the "best" choice is the first p left singular vectors of the matrix.

End of Extra Notes