CISC 271 Class 17

PCA – Matrix Algebra and Dimensionality Reduction

Text Correspondence: $\sim$

Main Concepts:

- *Scatter matrix*
- *PCA for dimensionality reduction*
- *Clustering of reduced data*

Sample Problem, Machine Inference: How can PCA reduce the number of variables in data?

We previously presented principal components analysis (PCA) by exploring the covariance of data. A data matrix A has a mean row-like matrix $\bar{A}$ and a zero-mean matrix M that we can re-state as

$$\begin{array}{rcl} \bar{A} & \stackrel{\text{def}}{=} & \frac{1}{m}\vec{1}^T A \\ M & \stackrel{\text{def}}{=} & A - \vec{1}\,\bar{A} \end{array} \tag{17.1}$$

Using the sample variance, the matrix M in Equation 17.1 can be used to define the covariance matrix as

$$B = \left(\frac{1}{m-1}\right) M^T M \tag{17.2}$$

The eigenvalues of the matrix B in Equation 17.2 are non-negative and can be interpreted as describing statistical relationships of columns of the data matrix A. A closely related matrix is the *scatter* matrix. This is a symmetric positive semidefinite matrix that we will define as

$$S \stackrel{\text{def}}{=} M^T M \tag{17.3}$$

Because the matrix S in Equation 17.3 is a scalar multiple of the matrix B in Equation 17.2, the matrix S and the matrix B have the same eigenvectors. The eigenvalues are also scalar multiples, that is, the j^{th} eigenvalue of S is $m-1$ times the j^{th} eigenvalue of B.

We can think of the scatter matrix S as the *weighted covariance*. The scatter matrix can be used in place of the covariance matrix when the number of observations are of concern, or when the interpretation of the eigenvalues of the covariance matrix are unaffected by the factor of $m-1$.

Henceforth, we will work with PCA using the scatter matrix B instead of using the covariance matrix B.

17.1 Scatter Matrix, SVD, and PCA

Using terminology that is common in PCA literature, we can say the PCA loading vectors $\vec{v}_j$ of the covariance matrix B are the same as the eigenvectors of the scatter matrix S.

Consider the SVD of the zero-mean matrix, which is $M = U\Sigma V^T$. We can write the scatter matrix as

$$\begin{aligned} S &= M^T M \\ &= [U\Sigma V^T]^T U\Sigma V^T \\ &= V\Sigma^T \Sigma V^T \\ &= V\Sigma^2 V^T \end{aligned} \tag{17.4}$$

As an aside, because the covariance matrix B is a scalar multiple of the scatter matrix S, we can use Equation 17.4 to write the spectral decomposition for PCA as

$$\begin{aligned} B &= \left(\frac{1}{m-1}\right) M^T M \\ &= V \left[\frac{\Sigma^2}{m-1}\right] V^T \end{aligned} \tag{17.5}$$

Equation 17.4 provides us with the spectral decomposition of the scatter matrix S. We can observe that the eigenvectors $\vec{v}_j$ are the right singular vectors of the data matrix M. The scatter matrix is symmetric and it is positive definite if and only if the zero-mean matrix M is full rank, that is, if and only if M has each singular value greater than zero. Because of Equation 17.5, these observation also apply to the covariance matrix B.

For PCA, we defined the first score vector as the product of the zero-mean matrix M and the first loading vector $\vec{v}_1$, that is, as $\vec{z}_1 = M\vec{v}_1$. The right singular matrix V of the zero-mean matrix M has orthonormal columns, which can be summarized as

$$\begin{aligned} \vec{v}_j^T \vec{v}_j &= 1 \\ \vec{v}_j^T \vec{v}_{i \neq j} &= 0 \end{aligned} \tag{17.6}$$

Using the properties of Equation 17.6, we have

$$V^T \vec{v}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \Rightarrow \quad \Sigma V^T \vec{v}_1 = \begin{bmatrix} \sigma_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{17.7}$$

The first score vector $\vec{z}_1$, using the SVD of the zero-mean matrix M and Equation 17.7, can be written as

$$\begin{aligned} \vec{z}_1 &= M\vec{v}_1 \\ &= U\Sigma V^T\vec{v}_1 \\ &= \sigma_1\vec{u}_1 \end{aligned} \tag{17.8}$$

Using the same reasoning of Equation 17.8 for each score, we can write the j^{th} score vector of the data matrix A as

$$\vec{z}_j = \sigma_j\vec{u}_j \tag{17.9}$$

17.2 PCA and Low-Rank Approximation

Recall, from a previous class, that the optimal approximation of a matrix is a truncation of the Eckart-Young series. This series, for the zero-mean matrix M that has rank r, is

$$\begin{aligned} M &= U\Sigma V^T \\ &= [\sigma_1\vec{u}_1]\vec{v}_1^T + [\sigma_2\vec{u}_2]\vec{v}_2^T + \cdots + [\sigma_r\vec{u}_r]\vec{v}_r^T \end{aligned} \tag{17.10}$$

Consider the first $p \leq r$ scores of the data in the original matrix A. These are the vectors $\vec{z}_j$ of Equation 17.9; we can gather these into a matrix Z_p. The first *p* eigenvectors of the scatter matrix S are the first *p* right singular vectors of the zero-mean matrix M, which we can gather into a matrix V_p. The *p* PCA scores and the *p* PCA loading vectors – which are the right singular vectors of M – are related by Equation 17.10 as

$$\begin{aligned} Z_p &= MV_p \\ &= \begin{bmatrix} \sigma_1\vec{u}_1 & \sigma_2\vec{u}_2 & \cdots & \sigma_p\vec{u}_p \end{bmatrix} \end{aligned} \tag{17.11}$$

An orthonormal basis for the column space of Z_p is, simply, to divide each column by the non-zero singular value σ_j. Each basis vector is $\vec{u}_j$, which is a left singular vector of the zero-mean matrix M; together, an orthonormal basis for the first *p* PCA scores are the first *p* columns of the left singular matrix U, which we can abbreviate as U_p. This leads us to a remarkable observation:

> *The first* p *PCA scores are an optimal approximation of a* p*-D vector space for the zero-mean data*

17.3 PCA for Dimensionality Reduction

PCA is commonly used to reduce the dimensionality of data. Dimensionality reduction is the process of transforming a linear problem that has n variables to a smaller problem that has $p < n$ variables. We have derived three equivalent methods to perform dimensionality reduction:

- Compute the first *p* scores of the data matrix A by using PCA, which is Z_p
- Compute the first *p* eigenvectors of the scatter matrix S as V_p and compute $U_p \Sigma_p = M V_p$
- Compute the SVD of the zero-mean matrix M and use the first *p* left singular vectors U_p

These equivalent methods of computing dimensionality reduction have the same effect. For the original *n*-D vector space of the zero-mean data, which we can abbreviate as $\mathbb{U}_n$, the methods find an orthogonal basis for a *p*-D vector space, which we can abbreviate as $\mathbb{U}_p$, with the important property that

$$p < n \textbf{ and } \mathbb{U}_p \subset \mathbb{U}_n$$

The latter property holds because the orthogonal left singular matrix U is an orthonormal basis for the column space of the zero-mean matrix M, so the first n columns U_n are a basis for $\mathbb{U}_n$.

Clustering algorithms have been found, from much empirical experience, to often perform poorly on high-dimensional data. A commonly used solution is to transform the data to a lower-dimensional vector space, and then to perform clustering by a method such as k-means.

17.4 Approximations and The Scree Plot

How can we select the rank p of an approximation? In practice, there is no universally accepted method. There may be guidelines within a domain of application but, in general, we usually examine the data and exercise human judgement.

A common method for determining a matrix approximation is the use of a *scree plot* [2]. Catell devised the name by a visual analogy to "... the straight line of rubble and boulders which forms at the pitch of sliding stability at the foot of a mountain".

In terms of the SVD, a scree plot has a horizontal axis, or ordinate, the is the integers i of Equation 16.20, ranging from 1 to either r or a suitable number less than r.

The vertical axis, or abscissa, is a normalized version of the singular value σ_j. Most commonly, the squares of the singular values is used because this is the *explained variance* s_j, defined as

$$s_i \stackrel{\text{def}}{=} \frac{\sigma_i^2}{\sum_{j=1}^{r} \sigma_j^2} \tag{17.12}$$

A value p for the approximate dimensionality of the vector space that contains the data can be found from a user-provided value θ as

$$p \stackrel{\text{def}}{=} \underset{k \leq r}{\operatorname{argmin}} \left(\theta \leq \sum_{j=1}^{k} s_j \right) \tag{17.13}$$

In Equation 17.13 we are selecting the smallest number of explained variances that add up to at least θ, starting with the largest s_j. The choice of θ depends on the application.

A simple example can be developed by generating a random matrix of integers. For one such matrix, which is $A \in \mathbb{R}^{20 \times 20}$ with $a_{ij} \leq 200$, the scree plot is shown in Figure 17.1. In this example the normalization was the sum of the squares of the singular values, from Equation 17.12. The scree plot suggests that the 20×20 matrix has a column space that can be approximated with 2 basis vectors, because the index 2 is where the scree plot begins to level.

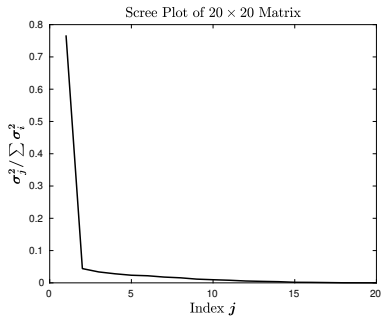

Figure 17.1: Scree plot of a random 20×20 integer-valued matrix. The index $p = 2$ would be chosen as the rank of approximation of this matrix.

The scree plot is one way that we can use the Eckart-Young theorem to select the number of basis vectors when we want to approximate the column space of data. These basis vectors are part of *principal component analysis*, which we will explore after we better understand vector spaces and projections.