Application to Data Compression

Suppose a satellite transmits a picture containing 1000×1000 pixels. If the color of each pixel is digitized, this information can be represented in a 1000×1000 matrix A.

Suppose we know an SVD

$$A = \sigma_1 \vec{u}_1 \vec{v}_1^T + \ldots + \sigma_r \vec{u}_r \vec{v}_r^T$$

Even if the rank r of the matrix A is large, most of the singular values will typically be very small (relatively to σ_1). If we neglect those, we get a good approximation $A \approx \sigma_1 \vec{u}_1 \vec{v}_1^T + \ldots + \sigma_s \vec{u}_s \vec{v}_s^T$, where s is much smaller than r.

For example, if we choose $s = 10$, we need to transmit only the 20 vectors $\sigma_1 \vec{u}_1, \ldots, \sigma_{10} \vec{u}_{10}$ and $\vec{v}_1, \ldots, \vec{v}_{10}$ in R^{1000}, that is, 20,000 numbers.

Application to Information Retrieval

Consider the problem of searching a database for documents. If there are m possible key words and a total of n documents. Then the database can be represented by a $m \times n$ matrix A.

Two of the main problems are polysemy (words having multiple meanings) and synonymy (multiple words having the same meaning).

If we think of our database as an approximation. Some of the entries may contain extraneous components due to polysemy, and some may miss including components because of synonymy.

Suppose it were possible to correct for these problems and come up with a perfect database matrix P. Let $E = A - P$, then $A = P + E$.

We can think of E as a matrix representing the errors.

Latent semantic indexing (LSI)

The idea of LSI is that the lower-rank matrix may still provide a good approximation to P and, may actually involve less error.

The lower-rank approximation can be obtained by truncating the outer product expansion of the singular value decomposition of A. This is equivalent to setting

$$\sigma_{s+1} = \sigma_{s+2} = \ldots = \sigma_n = 0$$

and then setting $A_s = U_s \Sigma_s V_s^T$, the compact form of the singular value decomposition.

Speedup

The matrix vector multiplication $A^T \vec{q}$ requires a total of mn scalar multiplications.

On the other hand, $A_s^T = V_s \Sigma_s U_s^T$, and the multiplication $A_s^T \vec{q} = V_s(\Sigma_s(U_s^T \vec{q}))$ requires a total of $s(m+n+1)$ scalar multiplications.

Reference

S. J. Leon, Linear algebra with applications, 6th Ed., Prentice Hall. 2002.

Applications to Statistics

`Matrix of observations`

An example of two-dimensional data is given by a set of weights and heights of N college students. Let X_j denote the observation vector in R^2 that lists the weight and height of the jth student. Then, the matrix of observation has the form

$$\begin{array}{cccc} \left[\begin{array}{cccc} w_1 & w_2 & \ldots & w_N \\ h_1 & h_2 & \ldots & h_N \end{array}\right] \\ \begin{array}{cccc} \uparrow & \uparrow & & \uparrow \\ X_1 & X_2 & \ldots & X_N \end{array} \end{array}$$

`Mean and Covariance`

To prepare for principle component analysis, let $\left[\begin{array}{ccc} X_1 & \ldots & X_N \end{array}\right]$ be a $p \times N$ matrix of observations. The sample mean, M, of the observation vectors is given by

$$M = \frac{1}{N}(X_1 + \ldots + X_N)$$

Let

$$\hat{X}_k = X_k - M$$

The columns of the $p \times N$ matrix

$$B = \left[\begin{array}{cccc} \hat{X}_1 & \hat{X}_2 & \dots & \hat{X}_N \end{array} \right]$$

have a zero sample mean, and B is said to be in `mean-deviation form`.

The `(sample) covariance matrix` is the $p \times N$ matrix S defined by

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

The entries s_{jj} is called the `variance` of x_j.
The `total variance` of the data is the sum of the variances on the diagonal of S, $totalvariance = trace(S)$.
The entries s_{ij} for $i \neq j$ is called the `covariance` of x_i and x_j.

Principle Component Analysis

Assume that the matrix $X = \begin{bmatrix} X_1 & \ldots & X_N \end{bmatrix}$ is already in mean-deviation form. The goal of principle component analysis is to find an orthogonal $p \times p$ matrix $P = \begin{bmatrix} u_1 & \ldots & u_p \end{bmatrix}$ that determines a change of variable, $X = PY$, or

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} u_1 & u_2 & \ldots & u_p \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix}$$

such that the new variables $y_1, y_2, \ldots, y_p$ are uncorrelated and are arranged in order of decreasing variance.

Let $S = \frac{1}{N-1} X X^T$ be the covariance matrix of X. Since the covariance matrix of $Y = \begin{bmatrix} Y_1 & \ldots & Y_N \end{bmatrix}$ is $\frac{1}{N-1} Y Y^T = \frac{1}{N-1} (P^T X)(P^T X)^T = P^T S P$. So the desired orthogonal matrix P is one that makes $P^T S P$ diagonal.

Let D be a diagonal matrix with the eigenvalues $\lambda_1, \ldots, \lambda_p$ of S on the diagonal, arranged that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, and let P be an orthogonal matrix whose columns are the corresponding unit eigenvectors $u_1, \ldots, u_p$. Then $P^T SP = D$ and $S = PDP^T$.

The unit eigenvectors $u_1, \ldots, u_p$ are called the `principle components` of the data. The first principle component u_1 determines the new variable y_1 in the following way. Let $c_1, \ldots, c_p$ be the entries in u_1. Since u_1^T is the first row of P^T, the equation $Y = P^T X$ shows that

$$y_1 = u_1^T X = c_1 x_1 + c_2 x_2 + \ldots + c_p x_p$$

Thus, y_1 is a linear combination of the original variables $x_1, x_2, \ldots, x_p$, using the entries in the eigenvector u_1 as weights.

Reducing the Dimension

Principle component analysis is potentially valuable for applications in which most of the variation in the data is due to variations in only a few of the new variables, $y_1, y_2, \ldots, y_p$.

The variance of y_j is λ_j, and the quotient $\lambda_j/trace(S)$ measures the fraction of the total variance that is captured by y_j.

Reference

D. C. Lay, Linear algebra and its applications, 2nd Ed. Addison-Wesley, 2000.