Principal Component Analysis

Yanjun Yan

2/12/04
ELE791, Spring 2004

Outline

• Introduction of PCA
• Mathematical basis
• Calculation of PCA
• Applications

What is PCA?

Principal Component Analysis, originally developed by Hotelling (1933), involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component is uncorrelated to former components and accounts for as much of the remaining variability as possible.

Objectives of PCA

• To discover or to reduce the dimensionality of the data set.
• To identify new meaningful underlying variables.
Motivation
1. Given observations \( \{x_1, x_2, \ldots, x_n\} \), the \( x \)'s will ordinarily be correlated. Is there a fundamental uncorrelated set, perhaps fewer in number than the \( x \)'s, which determine the values that \( x \)'s will take?
2. If \( y_1, y_2, \ldots, y_n \) are such variables, we shall have a set of relations of the form
   \[ x_i = f(y_1, y_2, \ldots) \quad (i = 1, 2, \ldots) \]

Assumptions
3. Consider only normally distributed systems of components having zero mean and unit variances.
   \[ E_{y_i} = 0 \]
   \[ E_{y_i y_j} = \delta_{ij} \] \hspace{1cm} (1)

Standard measures
4. In order to meet the assumption in 3, we can express the \( x \)'s in “standard measures”, by taking the deviation of each from its mean value and dividing its standard deviation. Thus we can obtain a set of quantities \( \{z_1, z_2, \ldots, z_n\} \) for which our formulas will be simpler. \( z_i = (x_i - \bar{x}) / \text{var}(x) \)

Linear Transforms
5. Confining ourselves to the case in which the functions \( f \) are linear, then
   \[ z_i = \sum a_{ij} y_j \] \hspace{1cm} (2)
   There might be less \( y \)'s than \( z \)'s if there’re fewer components than samples, and above formula includes this special case when \( a_{ij} = 0 \). However we’ll first assume that this is not the case and the determinant of \( A \) is not zero to see how we should determine \( a_{ij} \).

Cofactor
6. Let \( A_{ij} \) denote the cofactor of \( a_{ij} \) in \( A \) divided by determinant of \( A \). Then
   \[ \sum a_{ij} A_{ij} = \delta_{ij} \quad , \quad \sum a_{ij} A_{ij} = \delta_{ij} \] \hspace{1cm} (3)

7. Solve (2) for the \( y \)'s by multiplying both sides by \( A^{-1} \), summing with respect to \( j \) from 1 to \( n \), and using (3). Since \( \delta_{ij} \) is a sum consisting of terms which vanish except \( Y_i \), therefore
   \[ \sum a_{ij} z_i = \sum a_{ij} A_{ij} y_j = \sum \delta_{ij} y_j = y_i \] \hspace{1cm} (4)

Correlation
8. Let \( r_{ik} \) be the correlation between \( z_i \) and \( z_k \), equal to unity if \( i = k \);
9. Substitute the value for \( z_i \) given by (2). With the help of (1), we then obtain:
   \[ r_{ik} = E_{z_i z_k} = k \sum_{j=1}^{n} a_{ij} a_{jk} E_{y_i y_k} \]
   \[ = \sum_{j=1}^{n} a_{ij} a_{jk} = \sum_{j=1}^{n} a_{ik} a_{jk} \] \hspace{1cm} (5)
Rigid Rotation

10. Since $r_{ij} = r_{ji}$, the number of equations (5) is only $\frac{1}{2}(n+1)$. They are therefore insufficient for determining the $n^2$ quantities $a_{ij}$ when the correlation between the samples are known. Thus systems of uncorrelated components $\gamma$ may be chosen, consistently with the observed correlations, in $\frac{1}{2}(n(n-1))$ degrees of freedom of a rigid rotation in a space of $n$ dimensions.

Indeterminateness

11. The number $n$ of unknown $a_{ij}$ may be reduced by supposing that there are fewer than $n$ components, which amounts to setting some of the $a_{ij}$ equal to zero. Warning: If arbitrarily specialize the $a_{ij}$, the number of components possibly even exceeds the number of samples.

Goal in choosing the components

12. Begin with a component $\lambda_1$ whose contributions to the variances of the $x$'s have as great a total as possible; then we next take a component $\lambda_2$, uncorrelated with $\lambda_1$, whose contribution to the residual variance is as great as possible; and then we proceed in this way to determine the components, not exceeding $n$ in number, and perhaps neglecting those whose contributions to the total variance are small. This is called the method of principal components.

Picturization

13. If $z_1, z_2, \ldots, z_n$ be taken as rectangular coordinates in $n$ dimensions, each point represents a possible individual. If, as we assume, the population is normally distributed, the loci of uniform density are concentric, similar ellipsoids. The method of principal components is equivalent to choosing a set of coordinate axes coinciding with the principal axes of these ellipsoids.

Principal Components

14. Now that the set of $x$'s is capable of transformations such as changes of units and other linear transformations, the ellipsoids may be squeezed and stretched in any way. Thus for each $x_i$, there exits a unit of measure of unique importance. In other words, a metric – a definition of distance – must be assumed in the $n$-dimensional space. For different applications, different metrics would be suitable.
Mathematical Setup

15. Given
\[ z_i = \sum_{j=1}^{p} a_{ij} y_j \]  \hspace{1cm} (2)

The variance of \( z_i \) may be written as
\[ \text{var}(z_i) = E(z_i^2) = E\left( \sum_{j=1}^{p} a_{ij} y_j \right)^2 = \sum_{j=1}^{p} a_{ij}^2 \gamma_j \]

The first term in the sum is the contribution of \( \gamma_j \) to the variance of \( z_i \). The sum of the contributions of \( \gamma_j \) to the variances of all the \( z \)'s is
\[ S = \sum_{i=1}^{n} a_i^2 \]  \hspace{1cm} (6)

Maximization

16. We want to maximize (6) subject to (5)
\[ z_i = \sum_{j=1}^{p} a_{ij} \gamma_j \]
To this end we write
\[ 2T = S - \sum_{i=1}^{n} \sum_{j=1}^{l} \lambda_j a_{ij} - r_{ij} \]
where the \( \lambda_j = \lambda_k \) are Lagrange multipliers. Set
\[ \frac{\partial T}{\partial a_{ij}} = a_{ij} - \sum_{j=1}^{l} \lambda_j a_{ij} = 0 \]  \hspace{1cm} (7)
\[ \frac{\partial T}{\partial a_{ik}} = -\sum_{j=1}^{l} \lambda_j a_{ik} = 0 \]  \hspace{1cm} (8)

Final Formula for 1st PC

17. From (8) we can utilize system rank information to find an expression of \( \lambda_n \), then substitute it back to (7), we can get a bunch of equations:
\[ (1-k)a_{i1} + r_{i1} a_{i2} + \cdots + r_{n1} a_{ni} = 0 \]
\[ r_{i1} a_{i1} + (1-k)a_{i2} + \cdots + r_{n1} a_{ni} = 0 \]
\[ \vdots \]
\[ r_{ni} a_{i1} + r_{ni} a_{i2} + \cdots + (1-k)a_{ni} = 0 \]

Final Formula for succeeding PCs

18. Former formula is already very familiar to us:
\[ a \text{ be the vector of } [a_{11}, a_{21}, \ldots, a_{ni}]^T \]
\[ R \text{ be the covariance matrix with 1's at diagonal} \]
\[ (R - kI)a = 0 \]
Then \( a \) is the eigenvector of \( R \), and \( k \) is the eigenvalue corresponding to \( a \).

Final Formula for 1st PC (Con’t)

19. Next we need to find a component making a maximum contribution to the residual portion of the variance.
Change the second subscript in (6), (7) and (8) from 1 to 2, 3, \ldots. The arguments and procedure are virtually the same as before.

Meaning of \( k \)

20. For clarification, set the \( k \) for the first PC as \( \hat{k} \), and the succeeding \( k \)'s as \( \hat{k}_2, \hat{k}_3, \ldots \), then it can be shown that
\[ \gamma_1^2 + \gamma_2^2 + \cdots + \gamma_k^2 = \text{constant} \]
The \( \gamma \)'s are the PCs in original coordinates.
\( k \)'s are the length of the axes of the ellipsoids. If, instead of the \( z \)'s, the \( y \)'s be taken as rectangular coordinates, the ellipsoids are squeezed and stretched into spheres.
Outline

• Introduction of PCA
• Mathematical basis
• Calculation of PCA
• Applications

Basic Calculation of PCA

1. Eigendecompose the signal’s true (or estimated) covariance matrix.
2. Sort the eigenvalues from big to small, and sort the eigenvectors correspondingly.
3. According to the application, select several most significant eigenvectors, then use the weighting in the eigenvectors to linearly combine the raw data to get corresponding principal components.

SVD

The singular value decomposition (SVD) of the \( N \times p \) matrix \( X \) has the form

\[
X = UDV^T.
\]

Eigendecomposition

The sample covariance matrix of \( X \) is given by \( S = X^TX/N \), then

\[
X^TX = VDV^T
\]

Which is the eigendecomposition of \( X^TX \) (and of \( S \), up to a factor \( N \)). The eigenvectors are called principal components directions (or Karhunen-Loeve directions) of \( X \).

The first principal component direction \( v_1 \) has the property that \( x_1 = Xv_1 \) has the largest sample variance amongst all normalized linear combinations of the columns of \( X \).

Reconstruction from PCs

• For a given raw data sample, do the dot product with PCs to construct a reconstruction weighting vector.
• Given all PCs, the raw data sample can be reconstructed by linearly combine the PCs with the reconstruction weights.

My experience

Given a sample matrix, the estimation error in the covariance matrix may accumulate in the calculation of the principal components. The direct SVD of the sample matrix can yield better principal components at least from the image reconstruction point of view.
Outline

• Introduction of PCA
• Mathematical basis
• Calculation of PCA
• Applications

Applications

• Fisher classification
• Remote sensing multiband information extraction
• Optical character recognition (OCR) or handwriting recognition
• Face recognition (Eigenface)
• Cancer diagnosis
• ……..

eg.1 Handwriting Raw Data

PCA representation

\( \nu_1 \) (horizontal movement) mainly accounts for the lengthening of the lower tail of the three, while \( \nu_2 \) (vertical movement) accounts for character thickness.

Principal Component Space

eg.2 Remote Sensing

Vectorized Pixels

\[ \begin{align*}
\text{Raw data matrix} & \quad \text{Vectorized Pixels} \\
\text{Bands} & \quad \text{PCs matrix}
\end{align*} \]

PCs represent 'New Bands' made up of correlated combinations of the original bands. \( (M \leq N) \)
Vectorized Pixels

Suppose $p > N$, we can get $p$ eigenvectors with length $N$

1st eigenvector: Set $\sum = its\ eigenvalue, its\ elements\ are\ correlation\ coefficients$

False color image

Common Problems

- Image dimension mismatch. Appropriate compression or interpolation is needed.
- Calculation complexity and data scarcity. For $N \times p$ matrix $X$, the sample covariance requires $O(Np^2)$ operations; the snap shot algorithm requires $O(N^2)$ operations; EM algorithm requires $O(Np^2)$ operations ($r$ is the number of leading eigenvectors).

Common Problems (con’t)

- The discrimination power of principal components is not monotonically decreasing

Common Problems (con’t)

- One of the assumptions of the method is a linearity of correlation between samples. This is rarely met.
Similar yet Different Techniques

• Canonical Analysis (CA). Whereas PCA uses all pixels regardless of identity or class to derive the components, in CA one limits the pixels involved to those associated with pre-identified features/classes. This requires that these features can be recognized (by photointerpretation) in an image display (single band or color composite) in one to several areas within the scene. These pixels are "blocked out" as training sites. Their multiband values (within the site areas) are then processed in the manner of PCA. This selective approach is designed to optimize recognition and location of the same features elsewhere in the scene.

Similar yet Different Techniques (Con’t)

• If the function \( f_i \) is not linear, then we can get Nonlinear Component Analysis (NCA).
• Independent Component Analysis (ICA): ICA is a particular rotation method of factor analysis to make the bases statistically independent rather than uncorrelated.

Bibliography

This time:

Next time:

Any questions for me?

Q: If you are allowed to use only one word, which word would you use to characterize PCA?

Thanks!