Dimensionality Reduction: PCA

Nicholas Ruozzi
University of Texas at Dallas
Eigenvalues

• \( \lambda \) is an eigenvalue of a matrix \( A \in \mathbb{R}^{n \times n} \) if the linear system \( Ax = \lambda x \) has at least one non-zero solution

  – If \( Ax = \lambda x \) we say that \( \lambda \) is an eigenvalue of \( A \) with corresponding eigenvector \( x \)

  – Could be multiple eigenvectors for the same \( \lambda \)
Eigenvalues of Symmetric Matrices

- If $A \in \mathbb{R}^{n \times n}$ is symmetric, then it has $n$ linearly independent eigenvectors $v_1, \ldots, v_n$ corresponding to $n$ real eigenvalues.
  
  - Moreover, it has $n$ linearly independent orthonormal eigenvectors:
    
    - $v_i^T v_j = 0$ for all $i \neq j$
    - $v_i^T v_i = 1$ for all $i$
Eigenvalues of Symmetric Matrices

- If $A \in \mathbb{R}^{n \times n}$ is symmetric, then it has $n$ linearly independent eigenvectors $v_1, \ldots, v_n$ corresponding to $n$ real eigenvalues.

- A symmetric matrix is **positive definite** if and only if all of its eigenvalues are positive.
Example

- The 2x2 identity matrix has all of its eigenvalues equal to 1 with orthonormal eigenvectors $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

- The matrix $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ has eigenvalues 0 and 2 with orthonormal eigenvectors $\begin{bmatrix} -1 \\ \sqrt{2} \end{bmatrix}$ and $\begin{bmatrix} 1 \\ \sqrt{2} \end{bmatrix}$.
Eigenvalues

- Suppose $A \in \mathbb{R}^{n \times n}$ is symmetric

- Any $x \in \mathbb{R}^n$ can be written as $x = \sum_{i=1}^{n} c_i v_i$ where $v_1, \ldots, v_n$ are the eigenvectors of $A$

\[
\begin{align*}
- \quad Ax &= \sum_{i=1}^{n} \lambda_i c_i v_i \\
- \quad A^2 x &= \sum_{i=1}^{n} \lambda_i^2 c_i v_i \\
\vdots \\
- \quad A^t x &= \sum_{i=1}^{n} \lambda_i^t c_i v_i
\end{align*}
\]
Eigenvalues

• Suppose $A \in \mathbb{R}^{n \times n}$ is symmetric

• Any $x \in \mathbb{R}^n$ can be written as $x = \sum_{i=1}^{n} c_i v_i$ where $v_1, \ldots, v_n$ are the eigenvectors of $A$

$$-c_i = v_i^T x,$$ this is the projection of $x$ along the line given by $v_i$ (assuming that $v_i$ is a unit vector)
Eigenvalues of Symmetric Matrices

- Let $Q \in \mathbb{R}^{n \times n}$ be the matrix whose $i^{th}$ column is $v_i$ and $D \in \mathbb{R}^{n \times n}$ be the diagonal matrix such that $D_{ii} = \lambda_i$

  - $Ax = QDQ^T x$

  - Can throw away some eigenvectors to approximate this quantity

- For example, let $Q_k$ be the matrix formed by keeping only the top $k$ eigenvectors and $D_k$ be the diagonal matrix whose diagonal consists of the top $k$ eigenvalues
Frobenius Norm

- The Frobenius norm is a matrix norm written as

\[ \|A\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} |A_{ij}|^2} \]

- \( Q_k D_k Q_k^T \) is the best rank \( k \) approximation of the matrix symmetric matrix \( A \) with respect to the Frobenius norm.
Principal Component Analysis

• Given a collection of data points sampled from some distribution $x_1, \ldots, x_p \in \mathbb{R}^n$

  – Construct the matrix $X \in \mathbb{R}^{n \times p}$ whose $i^{th}$ column is $x_i$

• Want to reduce the dimensionality of the data while still maintaining a good approximation of the sample mean and variance
Principal Component Analysis

• Construct the matrix $\mathbf{W} \in \mathbb{R}^{n \times p}$ whose $i^{th}$ column is

$$x_i - \frac{\sum_j x_j}{p}$$

– This gives the data a zero mean

• The matrix $\mathbf{W}\mathbf{W}^T$ is the sample covariance matrix

– $\mathbf{W}\mathbf{W}^T$ is symmetric and positive semidefinite (simple proof later)
Principal Component Analysis

- PCA attempts to find a set of orthogonal vectors that best explain the variance of the sample covariance matrix.

  - From our previous discussion, these are exactly the eigenvectors of $WW^T$. 
PCA in Practice

• Forming the matrix $WW^T$ can require a lot of memory (especially if $n \gg p$)

  – Need a faster way to compute this without forming the matrix explicitly

  – Typical approach: use the singular value decomposition
Singular Value Decomposition (SVD)

- Every matrix $B \in \mathbb{R}^{n \times p}$ admits a decomposition of the form

$$B = U\Sigma V^T$$

- where $U \in \mathbb{R}^{n \times n}$ is an orthogonal matrix, $\Sigma \in \mathbb{R}^{n \times p}$ is non-negative diagonal matrix, and $V \in \mathbb{R}^{p \times p}$ is an orthogonal matrix

- A matrix $C \in \mathbb{R}^{m \times m}$ is orthogonal if $C^T = C^{-1}$. Equivalently, the rows and columns of $C$ are orthonormal vectors
Singular Value Decomposition (SVD)

- Every matrix \( B \in \mathbb{R}^{n \times p} \) admits a decomposition of the form

\[
B = U \Sigma V^T
\]

- where \( U \in \mathbb{R}^{n \times n} \) is an orthogonal matrix, \( \Sigma \in \mathbb{R}^{n \times p} \) is non-negative diagonal matrix, and \( V \in \mathbb{R}^{p \times p} \) is an orthogonal matrix

- A matrix \( C \in \mathbb{R}^{m \times m} \) is orthogonal if \( C^T = C^{-1} \). Equivalently, the rows and columns of \( C \) are orthonormal vectors
SVD and PCA

- Returning to PCA

  - Let $W = U\Sigma V^T$ be the SVD of $W$

  - $WW^T = U\Sigma V^T V \Sigma^T U^T = U\Sigma \Sigma^T U^T$

  - $U$ is then the matrix of eigenvectors of $WW^T$

  - If we can compute the SVD of $W$, then we don't need to form the matrix $WW^T$
SVD and PCA

• For any matrix $A$, $AA^T$ is symmetric and positive semidefinite
  
  – Let $A = UΣV^T$ be the SVD of $A$
  
  – $AA^T = UΣV^TVΣ^TUT = UΣΣ^TU^T$
  
  – $U$ is then the matrix of eigenvectors of $AA^T$
  
  – The eigenvalues of $AA^T$ are all non-negative because $ΣΣ^T = Σ^2$ which are the square of the singular values of $A$
An Example: “Eigenfaces”

• Let’s suppose that our data is a collection of images of the faces of individuals
An Example: “Eigenfaces”

• Let’s suppose that our data is a collection of images of the faces of individuals

  – The goal is, given the "training data", to correctly label unseen images

  – Let’s suppose that each image is an $s \times s$ array of pixels: $x_i \in \mathbb{R}^n, n = s^2$

  – As before, construct the matrix $W \in \mathbb{R}^{n \times p}$ whose $i^{th}$ column is $x_i - \sum_j \frac{x_j}{m}$
An Example: “Eigenfaces”

- Forming the matrix $WW^T$ requires a lot of memory
  - $s = 256$ means $WW^T$ is $65536 \times 65536$
  - Need a faster way to compute this without forming the matrix explicitly
  - Could use the singular value decomposition
An Example: “Eigenfaces”

- A different approach when \( p \ll n \)
  - Compute the eigenvectors of \( A^T A \) (this is an \( p \times p \) matrix)
  - Let \( v \) be an eigenvector of \( A^T A \) with eigenvalue \( \lambda \)
  - \( AA^T Av = \lambda Av \)
  - This means that \( Av \) is an eigenvector of \( AA^T \) with eigenvalue \( \lambda \) (or 0)
  - Save the top \( k \) eigenvectors - called eigenfaces in this example
An Example: “Eigenfaces”

• The data in the matrix is “training data”
  – Given a new image, we’d like to determine which, if any, member of the data set that it belongs to

• Step 1: Compute the projection of the recentered image to classify onto each of the $k$ eigenvectors
  – This gives us a vector of weights $c_1, \ldots, c_k$
An Example: “Eigenfaces”

• The data in the matrix is “training data”
  – Given a new image, we’d like to determine which, if any, member of the data set that it belongs to

• Step 2: Determine if the input image is close to one of the faces in the data set
  – If the distance between the input and it's approximation is too large, then the input is likely not a face
An Example: “Eigenfaces”

- The data in the matrix is “training data”
  - Given a new image, we’d like to determine which, if any, member of the data set that it belongs to

- Step 3: Find the person in the training data that is closest to the new input
  - Replace each group of training images by its average
  - Compute the distance to the $i^{th}$ average $\|c - a^i\|$ where $a^i$ are the coefficients of the average face for person $i$