EECS 275 Matrix Computation

Ming-Hsuan Yang

Electrical Engineering and Computer Science
University of California at Merced
Merced, CA 95344
http://faculty.ucmerced.edu/mhyang

Lecture 6
Overview

- Orthogonal projection, distance between subspaces
- Principal component analysis
Reading

- Chapter 6 of *Numerical Linear Algebra* by Llyod Trefethen and David Bau
- Chapter 2 of *Matrix Computations* by Gene Golub and Charles Van Loan
- Chapter 5 of *Matrix Analysis and Applied Linear Algebra* by Carl Meyer
Orthogonal projection

- Let $S \subset \mathbb{R}^n$ be a subspace, $P \in \mathbb{R}^{n \times n}$ is the orthogonal projection (i.e., projector) onto $S$ if $\text{ran}(P) = S$, $P^2 = P$, and $P^\top = P$.

- Mathematically, we have $y = Px$ for some $x$, then

$$Py = P^2x = Px = y$$

- Example, in $\mathbb{R}^3$

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad P \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x \\ y \\ 0 \end{bmatrix}, \quad \text{and} \quad P^2 \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x \\ y \\ 0 \end{bmatrix}$$

- For orthogonal projection,

$$P(Px - x) = P^2x - Px = P(I - P)x = 0$$

which means $Px - x \in \text{null}(P)$

- If $x \in \mathbb{R}^n$, then $Px \in S$ and $(I - P)x \in S^\perp$.
Orthogonal projection

- If \( P \) is a projector, \( I - P \) is also a projector, and
  \[
  \| I - P \|_2^2 = I - 2P + P^2 = I - P
  \]
  The matrix \( I - P \) is called complementary projector to \( P \)
- \( I - P \) projects to the null space of \( P \), i.e.,
  \[
  \text{ran}(I - P) = \text{null}(P)
  \]
  and, since \( P = I - (I - P) \), we have
  \[
  \text{null}(I - P) = \text{ran}(P)
  \]
  and \( \text{ran}(P) \cap \text{null}(P) = \{0\} \)
- If \( P_1 \) and \( P_2 \) are orthogonal projections, then for any \( z \in \mathbb{R}^n \), we have
  \[
  \| (P_1 - P_2)z \|_2^2 = (P_1 z)^\top (I - P_2) z + (P_2 z)^\top (I - P_1) z
  \]
- If \( \text{ran}(P_1) = \text{ran}(P_2) = S \), then the right hand side of the above equation is zero, i.e., the orthogonal projection for a subspace is unique
Orthogonal projection and SVD

- If the columns of $V = [v_1, \ldots, v_k]$ are an orthonormal basis for a subspace $S$, then it is easy to show that $P = V V^\top$ is the unique orthogonal projection onto $S$.

- If $v \in \mathbb{R}^n$, then $P = \frac{vv^\top}{v^\top v}$ is the orthogonal projection onto $S = \text{span} \{ \{v\} \}$.

- Let $A = U \Sigma V^\top \in \mathbb{R}^{m \times n}$ and $\text{rank}(A) = r$, we have the $U$ and $V$ partitionings

  $$U = \begin{bmatrix} U_r & \tilde{U} \end{bmatrix}, \quad V = \begin{bmatrix} V_r & \tilde{V} \end{bmatrix},$$

  with

  $U_r U_r^\top = \text{projection onto } \text{ran}(A)$

  $\tilde{U}_r \tilde{U}_r^\top = \text{projection onto } \text{ran}(A)^\perp = \text{null}(A^\top)$

  $V_r V_r^\top = \text{projection onto } \text{null}(A)^\perp = \text{ran}(A^\top)$

  $\tilde{V}_r \tilde{V}_r^\top = \text{projection onto } \text{null}(A)$
Distances between subspaces

- Let $S_1$ and $S_2$ be subspaces of $\mathbb{R}^n$ and $\dim(S_1) = \dim(S_2)$, we define the distance between two spaces by

$$\text{dist}(S_1, S_2) = \|P_1 - P_2\|_2$$

where $P_i$ is the orthogonal projection onto $S_i$.

- The distance between a pair of subspaces can be characterized in terms of the blocks of a certain orthogonal matrix.

Theorem

Suppose

$$W = \begin{bmatrix} W_1 & W_2 \end{bmatrix} \quad Z = \begin{bmatrix} Z_1 & Z_2 \end{bmatrix}$$

are $n$-by-$n$ orthogonal matrices. If $S_1 = \text{ran}(W_1)$, and $S_2 = \text{ran}(Z_1)$, then

$$\text{dist}(S_1, S_2) = \|W_1^\top Z_2\|_2 = \|Z_1^\top W_2\|_2$$

See Golub and Van Loan for proof.
Distance between subspaces in $\mathbb{R}^n$

- If $S_1$ and $S_2$ are subspaces in $\mathbb{R}^n$ with the same dimension, then
  $$0 \leq \text{dist}(S_1, S_2) \leq 1$$

- The distance is zero if $S_1 = S_2$ and one if $S_1 \cap S_2^\perp \neq \{0\}$
Symmetric matrices

- Consider real, symmetric matrices, $A^\top = A$,
  - Hessian matrix (second order partial derivatives of a function):
    \[
    y = f(x + \Delta x) \approx f(x) + J(x) \Delta x + \frac{1}{2} \Delta x^\top H(x) \Delta x
    \]
    where $J$ is the Jacobian matrix
  - covariance matrix for Gaussian distribution
- The inverse is also symmetric: $(A^{-1})^\top = A^{-1}$
- Eigenvector equation for a symmetric matrix
  \[
  Au_k = \lambda_k u_k
  \]
  which can be written as
  \[
  AU = DU, \text{ or } (A - D)U = 0
  \]
  where $D$ is a diagonal matrix whose elements are eigenvalues
  \[
  D = \begin{bmatrix}
  \lambda_1 \\
  \vdots \\
  \lambda_m
  \end{bmatrix}
  \]
  and $U$ is matrix whose columns are eigenvectors $u_k$
Eigenvectors for symmetric matrices

- The eigenvectors can be computed from determinant $|A - D| = 0$
- Eigenvectors can be chosen to form an orthonormal basis as follows
- For a pair of eigenvectors $\mathbf{u}_j$ and $\mathbf{u}_k$, it follows

\[
\begin{align*}
\mathbf{u}_j^\top A \mathbf{u}_k &= \lambda_k \mathbf{u}_j^\top \mathbf{u}_k \\
\mathbf{u}_k^\top A \mathbf{u}_j &= \lambda_j \mathbf{u}_k^\top \mathbf{u}_j
\end{align*}
\]

and since $A$ is symmetric, we have

\[
(\lambda_k - \lambda_j) \mathbf{u}_k^\top \mathbf{u}_j = 0
\]

- For $\lambda_k \neq \lambda_j$, the eigenvectors must be orthogonal
- Note for any $\mathbf{u}_k$ with eigenvalue $\lambda_k$, $\beta \mathbf{u}_k$ is also an eigenvector for non-zero $\beta$ with the same eigenvalue
- Can be used to normalize the eigenvectors to unit norm so that

\[
\mathbf{u}_k^\top \mathbf{u}_j = \delta_{kj}
\]
Symmetric matrices and diagonalization

- Since $A u_k = \lambda_k u_k$, multiply $A^{-1}$ and we obtain

$$A^{-1} u_k = \lambda_k^{-1} u_k$$

so $A^{-1}$ has the same eigenvectors as $A$ but with reciprocal eigenvalues

- For symmetric matrix $A$, $AU = DU$ and $U^\top U = I$, $U = [u_1, \ldots, u_m]$, $A$ can be diagonalized

$$U^\top AU = D$$

- For symmetric matrix $A$, the SVD of $A = U \Sigma U^\top$

- Recall $U$, $V$ are left and right singular vectors

$$\begin{align*}
(AA^\top) U &= \Sigma U \\
(A^\top A) V &= \Sigma V
\end{align*}$$

Since $A$ is symmetric, $U = V$, and $A = U \Sigma U^\top$
Principal component analysis (PCA)

- Arguably the most popular dimensionality reduction algorithm
- Curse of dimensionality
- Widely used in computer vision, machine learning and pattern recognition
- Can be derived from several perspectives:
  - Minimize reconstruction error: Karhunen-Loeve transform
  - Decorrelation: Hotelling transform
  - Maximize the variance of the projected samples (i.e., preserve as much energy as possible)
- Unsupervised learning
- Linear transform
- Second order statistics
- Recall from SVD we have $A = U \Sigma V^\top$, and thus project samples on the subspace spanned by $U$ can be computed by

$$U^\top A = \Sigma V^\top$$
Principal component analysis

- Given a set of \( n \) data points \( \mathbf{x} \in \mathbb{R}^m \), we would like to project each \( \mathbf{x}^{(k)} \) onto a \( d \)-dimensional subspace \( \mathbf{z}^{(k)} = [z_1, \ldots, z_d] \in \mathbb{R}^d \), \( d < m \), so that

\[
\mathbf{x} = \sum_{i=1}^{d} z_i \mathbf{u}_i
\]

where the vectors \( \mathbf{u}_i \) satisfy the orthonormality relation

\[
\mathbf{u}_i^\top \mathbf{u}_j = \delta_{ij}
\]

in which \( \delta_{ij} \) is the Kronecker delta. Thus,

\[
z_i = \mathbf{u}_i^\top \mathbf{x}
\]

- Now we have only a subset \( d < m \) of the basis vector \( \mathbf{u}_i \). The remaining coefficients will be replaced by constants \( b_i \) so that each vector \( \mathbf{x} \) is approximated by \( \mathbf{x} \) can be approximated by

\[
\hat{\mathbf{x}} = \sum_{i=1}^{d} z_i \mathbf{u}_i + \sum_{i=d+1}^{m} b_i \mathbf{u}_i
\]
Principal component analysis (cont’d)

- Dimensionality reduction: $\mathbf{x}$ has $m$ degree of freedom and $\mathbf{z}$ has $d$ degree of freedom, $d < m$
- For each $\mathbf{x}^{(k)}$, the error introduced by the dimensionality reduction is

$$
\mathbf{x}^{(k)} - \tilde{\mathbf{x}}^{(k)} = \sum_{i=d+1}^{m} (z_i^{(k)} - b_i) \mathbf{u}_i
$$

and we want to find the basis vector $\mathbf{u}_i$, the coefficients $b_i$, and the values $z_i$ with minimum error in $\ell_2$-norm
- For the whole data set, with orthonormality relation

$$
E_d = \frac{1}{2} \sum_{k=1}^{n} \| \mathbf{x}^{(k)} - \tilde{\mathbf{x}}^{(k)} \|^2 = \frac{1}{2} \sum_{k=1}^{n} \sum_{i=d+1}^{m} (z_i^{(k)} - b_i)^2
$$
Principal component analysis (cont’d)

- Take derivative of $E_d$ with respect to $b_i$ and set it to zero,
  \[
  b_i = \frac{1}{n} \sum_{k=1}^{n} z_i^{(k)} = \frac{1}{n} \sum_{k=1}^{n} u_i^\top x^{(k)} = u_i^\top \bar{x} \quad \text{where, } \bar{x} = \frac{1}{n} \sum_{k=1}^{n} x^{(k)}
  \]

- Plug it into the sum of square errors, $E_d$,
  \[
  E_d = \frac{1}{2} \sum_{i=d+1}^{m} \sum_{k=1}^{n} (u_i^\top (x^{(k)} - \bar{x}))^2
  = \frac{n}{2} \sum_{i=d+1}^{m} u_i^\top C u_i
  \]

  where $C$ is a covariance matrix

  \[
  C = \frac{1}{n} \sum_{k=1}^{n} (x^{(k)} - \bar{x})(x^{(k)} - \bar{x})^\top
  \]

- Minimizing $E_d$ with respect to $u_i$, we get
  \[
  C u_i = \lambda_i u_i
  \]

  i.e., the basis vectors $u_i$ are the eigenvectors of the covariance matrix $C$
Derivation

- Minimizing $E_d$ with respect to $u_i$,

$$E_d = \frac{1}{2} \sum_{i=d+1}^{m} \sum_{k=1}^{n} (u_i^T (x^{(k)} - \bar{x}))^2 = \frac{n}{2} \sum_{i=d+1}^{m} u_i^T C u_i$$

- Need some constraints to solve this optimization problem
- Impose orthonormal constraints among $u_i$
- Use Lagrange multipliers $\phi_{ij}$

$$\hat{E}_d = \frac{1}{2} \sum_{i=d+1}^{m} u_i C u_i^T - \frac{1}{2} \sum_{i=d+1}^{m} \sum_{j=d+1}^{m} \phi_{ij} (u_i^T u_j - \delta_{ij})$$

- Recall

$$\min f(x) \quad \text{s.t.} \quad g(x) = 0 \quad \Rightarrow L(x, \phi) = f(x) + \phi g(x)$$

- Example: $\min f(x_1, x_2) = x_1 x_2$ subject to $g(x_1, x_2) = x_1 + x_2 - 1 = 0$
Derivation (cont’d)

• In matrix form,

\[ \hat{E}_d = \frac{1}{2} \text{tr}\{U^\top C U\} - \frac{1}{2} \text{tr}\{M(U^\top U - I)\} \]

where \( M \) is a matrix with elements \( \phi_{ij} \), and \( U \) is a matrix whose columns are \( u_i \).

• Minimizing \( \hat{E}_d \) with respect to \( U \),

\[ (C + C^\top)U - U(M + M^\top) = 0 \]

• Note \( C \) is symmetric, \( M \) is symmetric since \( UU^\top \) is symmetric. Thus

\[ CU = UM \]

\[ U^\top CU = M \]

• Clearly one solution is to choose \( M \) to be diagonal so that the columns of \( U \) are eigenvectors of \( C \) and the diagonal elements of \( M \) are eigenvalues.
Derivation (cont’d)

- The eigenvector equation for $M$
  \[ M\Psi = \Psi \Lambda \]
  where $\Lambda$ is a diagonal matrix of eigenvalues and $\Psi$ is the matrix of eigenvectors

- $M$ is symmetric and $\Psi$ can be chosen to have orthonormal columns, i.e., $\Psi^\top \Psi = I$
  \[ \Lambda = \Psi^\top M \Psi \]

- Put together,
  \[ \Lambda = \Psi^\top U^\top C U \Psi \]
  \[ = (U \Psi)^\top C (U \Psi) \]
  \[ = \tilde{U}^\top C \tilde{U} \]
  where $\tilde{U} = U \Psi$, and
  \[ U = \tilde{U} \Psi^\top \]

- Another solution for $U^\top C U = M$ can be obtained from the particular solution $\tilde{U}$ by application of an orthogonal transformation given by $\Psi$
We note that $E_d$ is invariant under this orthogonal transformation

$$
E_d = \frac{1}{2} \text{tr}\{ U^\top C U \} = \frac{1}{2} \text{tr}\{ \Psi \tilde{U}^\top \tilde{C} \tilde{U} \Psi^\top \} = \frac{1}{2} \text{tr}\{ \tilde{U}^\top \tilde{C} \tilde{U} \}
$$

Recall the matrix 2-norm is invariant under orthogonal transformation

Since all of the possible solutions give the same minimum error $E_d$, we can choose whichever is most convenient

We thus choose the solutions given by $\tilde{U}$ (with unit norm) since this has columns which are the eigenvectors of $C$
Computing principal components from data

- Minimizing $E_d$ with respect to $u_i$, we get

$$Cu_i = \lambda_i u_i$$

i.e., the basis vectors $u_i$ are the eigenvectors of the covariance matrix $C$

- Consequently, the error of $E_d$ is

$$E_d = \frac{1}{2} \sum_{i=d+1}^{m} \lambda_i$$

In other words, the minimum error is reached by discarding the eigenvectors corresponding to the $m - d$ smallest eigenvalues

- Retain the eigenvectors corresponding to the largest eigenvalues
Computing principal components from data

- Project $\mathbf{x}^{(k)}$ onto these eigenvectors give the components of the transformed vector $\mathbf{z}^{(k)}$ in the $d$-dimensional space.

- Each two-dimensional data point is transformed to a single variable $z_1$ representing the projection of the data point onto the eigenvector $u_1$.

- Infer the structure (or reduce redundancy) inherent in high dimensional data.

- Parsimonious representation.

- Linear dimensionality algorithm based on sum-of-square-error criterion.

- Other criteria: covariance measure and population entropy.
Intrinsic dimensionality

- A data set in $m$ dimensions has intrinsic dimensionality equal to $m'$ if the data lies entirely within a $m'$-dimensional space.
- What is the intrinsic dimensionality of data?
- The intrinsic dimensionality may increase due to noise.
- PCA, as a linear approximation, has its limitation.
- How to determine the number of eigenvectors?
- Empirically determined based on reconstruction error (i.e., energy).