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 7

Overview

- Principal component analysis
- Karhunen-Loeve Transform
- Multivariate Gaussian
- Applications

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
- Chapter 2 of *Principal Component Analysis* by Ian Jolliffe

Karhunen-Loeve Transform

- Transform data into a new set of variables, the principal components (PC)
  - which are uncorrelated and ordered
  - so that the first few retain most of the variation
- Consider the first PC, $\mathbf{u}_1^\top \mathbf{x}$,

$$\mathbf{u}_1 = \arg\max_{\|\mathbf{u}\|=1} \mathrm{var}(\mathbf{u}^\top \mathbf{x}) = \arg\max_{\|\mathbf{u}\|=1} E[\mathbf{u}^\top \mathcal{C} \mathbf{u}]$$

- Solving constrained optimization with Lagrange multiplier

$$\mathbf{u}^\top \mathcal{C} \mathbf{u} - \lambda(\mathbf{u}^\top \mathbf{u} - 1)$$

- Take derivative with respect to $\mathbf{u}$

$$\mathcal{C}\mathbf{u} - \lambda \mathbf{u} = 0, \quad (\mathcal{C} - \lambda I)\mathbf{u} = 0$$

Thus, λ is an eigenvalue of $\mathcal{C}$ and $\mathbf{u}$ is the corresponding eigenvector

Karhunen-Loeve Transform (cont'd)

- To maximize $\text{var}(\mathbf{u}^\top \mathbf{x})$,

$$\mathbf{u}^\top \mathcal{C} \mathbf{u} = \mathbf{u}^\top \lambda \mathbf{u} = \lambda \mathbf{u}^\top \mathbf{u} = \lambda$$

  so $\mathbf{u}_1$ is the eigenvector corresponding to the largest eigenvalue of $\mathcal{C}$
- In general, the k-th PC of $\mathbf{x}$ is $\mathbf{u}_k^\top \mathbf{x}$ and $\text{var}(\mathbf{u}_k^\top \mathbf{x}) = \lambda_k$ where λ_k is the k-th largest eigenvalue
- The second PC, $\mathbf{u}_2 \mathbf{x}$ maximizes $\mathbf{u}_2 \mathcal{C} \mathbf{u}_2$ subject to being uncorrelated with $\mathbf{u}_1 \mathbf{x}$, i.e., $\text{cov}(\mathbf{u}_1^\top \mathbf{x}, \mathbf{u}_2^\top \mathbf{x}) = 0$

$$\text{cov}(\mathbf{u}_1^\top \mathbf{x}, \mathbf{u}_2^\top \mathbf{x}) = \mathbf{u}_1^\top \mathcal{C} \mathbf{u}_2 = \mathbf{u}_2^\top \mathcal{C} \mathbf{u}_1 = \mathbf{u}_2^\top \lambda_1 \mathbf{u}_1 = \lambda_1 \mathbf{u}_2^\top \mathbf{u}_1 = \lambda_1 \mathbf{u}_1^\top \mathbf{u}_2 = 0$$

- Solving constrained optimization problem with one of these constraints

$$\mathbf{u}_2^\top \mathcal{C} \mathbf{u}_2 - \lambda(\mathbf{u}_2^\top \mathbf{u}_2 - 1) - \phi \mathbf{u}_2^\top \mathbf{u}_1$$

  where λ, ϕ are Lagrange multipliers

Karhunen-Loeve Transform (cont'd)

- Take derivative with respect to $\mathbf{u}_2$

$$\mathcal{C}\mathbf{u}_2 - \lambda\mathbf{u}_2 - \phi\mathbf{u}_1 = 0$$

and multiply on the left by $\mathbf{u}_1$

$$\mathbf{u}_1^\top\mathcal{C}\mathbf{u}_2 - \lambda\mathbf{u}_1^\top\mathbf{u}_2 - \phi\mathbf{u}_1^\top\mathbf{u}_1 = 0$$

- Consequently $\phi = 0$ and $\mathcal{C}\mathbf{u}_2 = \lambda\mathbf{u}_2$
- Assuming that $\mathcal{C}$ does not have repeated eigenvalues, λ has to be the second largest eigenvalue to satisfy all the constraints

Karhunen-Loeve Transform and SVD

- Assuming $\mathbf{x}$ has zero mean, the principal component $\mathbf{u}_1$ is

$$\mathbf{u}_1 = \arg\max_{\|\mathbf{u}\|=1} \mathrm{var}(\mathbf{u}^\top \mathbf{x}) = \arg\max_{\|\mathbf{u}\|=1} E[(\mathbf{u}^\top \mathbf{x})^2]$$

- With the first $k-1$ component, the k-th component can be found by subtracting the first $k-1$ principal components from $\mathbf{x}$

$$\hat{\mathbf{x}}_{k-1} = \mathbf{x} - \sum_{i=1}^{k-1} \mathbf{u}_i \mathbf{u}_i^\top \mathbf{x}$$

  and find a principal component in

$$\mathbf{u}_k = \arg\max_{\|\mathbf{u}\|=1} E[(\mathbf{u}^\top \hat{\mathbf{x}}_{k-1})^2]$$

- The Karhunen-Loeve transform is therefore equivalent to finding the singular value decomposition of X

Karhunen-Loeve Transform and SVD (cont'd)

- A simpler way to compute the principal components
- Let $X = U\Sigma V^\top$, the projected data onto the subspace spanned by the first d singular vectors

$$Y = U_d^\top X = \Sigma_d V_d^\top$$

- The matrix U of singular vectors of X is equivalently the matrix U of eigenvectors of the covariance matrix $\mathcal{C}$

$$\mathcal{C} = XX^\top = U\Sigma\Sigma^\top U^\top$$

- The eigenvectors with the largest eigenvalues correspond to the dimensions that have the strongest correlation in the data set

Rayleigh quotient and eigenvectors

- The Rayleigh quotient for a real matrix M and vector $\mathbf{x}$ is

$$\rho(\mathbf{x}) = \frac{\mathbf{x}^\top M \mathbf{x}}{\mathbf{x}^\top \mathbf{x}}$$

and for eigenvector $\mathbf{u}$ w.r.t. covariance matrix $\mathcal{C}$

$$\rho(\mathbf{u}) = \frac{\mathbf{u}^\top \mathcal{C} \mathbf{u}}{\mathbf{u}^\top \mathbf{u}} = \lambda \frac{\mathbf{u}^\top \mathbf{u}}{\mathbf{u}^\top \mathbf{u}} = \lambda$$

- The eigenvectors $\mathbf{u}_i$ are the critical points of the Rayleigh quotient and their eigenvalues λ_i are the stationary values of $\rho(\mathbf{u})$
- To find the critical point of the Rayleigh quotient w.r.t. A

$$\begin{aligned} \rho(\mathbf{x}) &= \tfrac{\mathbf{x}^\top A \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \mathbf{x}^\top A \mathbf{x} \\ \text{s. t. } \|\mathbf{x}\|_2 &= 1 \end{aligned}$$

- The constrained optimization problem

$$L(\mathbf{x}, \lambda) = \mathbf{x}^\top A \mathbf{x} - \lambda(\mathbf{x}^\top \mathbf{x} - 1)$$

where λ is a Lagrange multiplier

Rayleigh quotient and eigenvectors (cont'd)

- Take derivative with respect to $\mathbf{x}$

$$\begin{aligned} 2A\mathbf{x} - 2\lambda\mathbf{x} &= 0 \\ A\mathbf{x} &= \lambda\mathbf{x} \end{aligned}$$

- Thus

$$\rho(\mathbf{x}) = \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} = \lambda\frac{\mathbf{x}^\top\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} = \lambda$$

- The eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of A are critical points of the Rayleigh quotient and their corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$ are the stationary values of $\rho(\mathbf{x})$
- Basis for PCA and canonical correlation analysis

Derivation using covariance matrix

- Let X be a m-dimensional vector with zero mean. We want to find a $m \times m$ orthonormal projection matrix P so that $Y = PX$ has a diagonal covariant matrix $\mathcal{C}_Y$ (i.e., Y is a vector with all its distinct components pairwise uncorrelated) and $P^\top = P$

$$\mathcal{C}_Y = E[YY^\top] = E[PX(PX)^\top] = PE[XX^\top]P^\top = P\mathcal{C}_X P^\top = P^\top \mathcal{C}_X P$$

- Therefore,

$$P\mathcal{C}_Y = PP^\top \mathcal{C}_X P = \mathcal{C}_X P$$

- Note $P = [\mathbf{p}_1, \ldots, \mathbf{p}_d]$ and $\mathcal{C}_Y = \mathrm{diag}\{\lambda_1, \ldots, \lambda_d\}$

$$[\lambda_1 \mathbf{p}_1, \lambda_2 \mathbf{p}_2, \ldots, \lambda_d \mathbf{p}_d] = [\mathcal{C}_X \mathbf{p}_1, \mathcal{C}_X \mathbf{p}_2, \ldots, \mathcal{C}_X \mathbf{p}_d]$$

  i.e., $\lambda_i \mathbf{p}_i = \mathcal{C}_X \mathbf{p}_i$, and $\mathbf{p}_i$ is an eigenvector of the covariance matrix, $\mathcal{C}_X$ of X

SVD and PCA

- Recall

$$\mathbf{x} = \sum_{i=1}^{m} z_i \mathbf{u}_i, \quad z_i = \mathbf{u}_i^\top \mathbf{x}, \quad , \text{ and } \quad \mathbf{u}_i^\top \mathbf{u}_j = \delta_{ij}$$

- Center the data points

$$X = [(\mathbf{x}^{(1)} - \bar{\mathbf{x}}) \ldots (\mathbf{x}^{(n)} - \bar{\mathbf{x}})]$$

  Covariance matrix

$$\mathcal{C} = XX^\top$$

- Singular value decomposition allows us to write X as

$$X = U\Sigma V^\top = \left[\begin{array}{ccc} \mathbf{u}_1 & \ldots & \mathbf{u}_n \end{array} \right] \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^\top \\ \vdots \\ \mathbf{v}_n^\top \end{bmatrix}$$

SVD and PCA (cont'd)

$$\begin{aligned} C &= \frac{1}{n}XX^\top \\ &= \frac{1}{n}U\Sigma V^\top(U\Sigma V^\top)^\top \\ &= \frac{1}{n}U\Sigma V^\top V\Sigma U^\top \\ &= \frac{1}{n}U\Sigma^2 U^\top \end{aligned}$$

- Therefore,

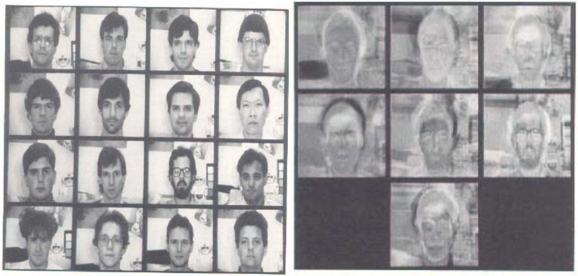
$$C\mathbf{u}_i = \frac{\sigma_i^2}{n}\mathbf{u}_i$$

- So, the columns U are eigenvectors and the eigenvalues are just $\lambda_i = \frac{\sigma_i^2}{n}$

Properties and limitations of PCA

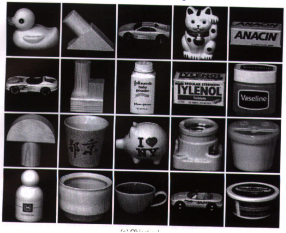
- Theoretically optimal subspace representation in terms of ℓ_2-norm
- Involves only rotation and scaling
- Unsupervised learning
- Unique solution
- Assumption:
  - data can be modeled linearly
  - data can be modeled with mean and covariance, i.e., Gaussian distribution
  - the large variances have important dynamics
  - ℓ_2-norm
- Nonlinear PCA, mixture of PCA, probabilistic PCA, mixture of probabilistic PCA, factor analysis, mixture of factor analyzers, sparse PCA, independent component analysis, Fisher linear discriminant, etc.

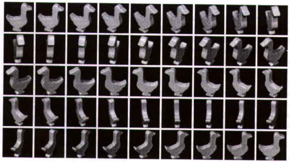
Eigenface [Turk and Pentland 1991]



- Collect a set of face images
- Normalize for contrast, scale and orientation
- Apply PCA to compute the first d eigenvectors (dubbed as Eigenface) that best accounts for data variance (i.e., facial structure)
- Compute the distance between the projected points for face recognition or detection

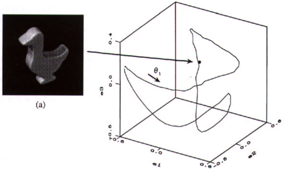
Appearance manifolds [Murase and Nayar 95]

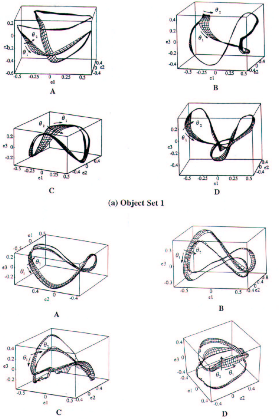




- The image variation of an object under different pose or is assumed to lie on a manifold
- For each object, collect images under different pose
- Construct a universal eigenspace from all the images
- For the set of images of of the same object, find the smoothly varying manifold in eigenspace, i.e., parametric eigenspace
- The manifolds of two objects may intersect, the intersection corresponds to poses of the two objects for which their images are very similar in appearance

Appearance manifolds [Murase and Nayar 95]





Gaussian distribution

- Univariate Gaussian

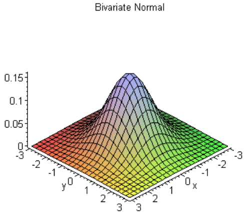
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

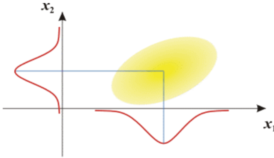
- Bivariate Gaussian

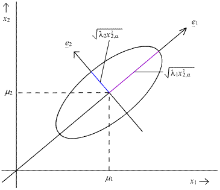
$$\begin{aligned} p(x,y) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \\ &\quad \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right)\right) \end{aligned}$$

where ρ is the correlation between x and y

$$\rho = \frac{\text{cov}(x,y)}{\sigma_x\sigma_y} = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x\sigma_y}$$







Multivariate Gaussian

- Multivariate Gaussian: $\mathbf{x} \in \mathbb{R}^d$

$$
\begin{aligned}
p(\mathbf{x}) &= \frac{1}{(2\pi)^{d/2}|\mathcal{C}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathcal{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \\
&= \frac{1}{(2\pi)^{d/2}|\mathcal{C}|^{1/2}} \exp\left(-\frac{1}{2}\Delta^2\right)
\end{aligned}
$$

  where

$$
\begin{aligned}
\boldsymbol{\mu} &= E[\mathbf{x}] \\
\mathcal{C} &= E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top] \\
\Delta &= C^{-1/2}(\mathbf{x}-\mathbf{u})
\end{aligned}
$$

  and Δ is called the Mahalanobis distance from $\mathbf{x}$ to μ
- The surfaces of constant probability density are hyperellipsoids on which Δ^2 is constant
- The principal axes of the hyperellipsoids are given by the eigenvectors $\mathbf{u}_i$ of $\mathcal{C}$ which satisfy

$$
\mathcal{C}\mathbf{u}_i = \lambda_i \mathbf{u}_i
$$

  and the corresponding eigenvalues λ_i give the variances along the respective principal directions