PCA

Lyle Ungar

Learning objectives

PCA as change of basis
PCA minimizes reconstruction error
PCA maximizes variance
PCA relation to eigenvalues/vectors

PCR: PCA for feature creation

Based in part on slides by Jia Li (PSU) and Barry Slaff (Upenn)

PCA

- **Express a vector *x* in terms of coefficients on an (orthogonal) basis vector (eigenvectors v_k)**

  $$\mathbf{x}_i = \Sigma_k \, z_{ik}\mathbf{v}_k$$

  - We can describe how well we approximate ***x*** in terms of the eigenvalues
- **PCA is used for dimensionality reduction**
  - visualization
  - semi-supervised learning
  - eigenfaces, eigenwords, eigengrasps

PCA

- Express a vector *x* in terms of coefficients on an (orthogonal) basis vector (eigenvectors v_k)

  $$\boldsymbol{x}_i = \sum_k z_{ik} \boldsymbol{v}_k$$

- Find z_{ik} by projection

  $$\boldsymbol{x}_i \cdot \boldsymbol{v}_j = \sum_k z_{ik} \boldsymbol{v}_k \cdot \boldsymbol{v}_j$$

  $$\boldsymbol{x}_i \cdot \boldsymbol{v}_j = z_{ij}$$

PCA

- **PCA can be viewed as**
  - minimizing distortion $||\mathbf{X} - \mathbf{Z}\mathbf{V}^T||_F$
    - Or the square of the above: $\Sigma_i ||\mathbf{x}_i - \Sigma_k z_{ik}\mathbf{v}_k||_2^2$
    - Note that either definition gives the same result
  - A rotation to a new coordinate system to maximize the variance in the new coordinates

- **Generally done by mean centering first**
  - Sometimes standardize

Nomenclature

$\mathbf{X} = \mathbf{Z}\mathbf{V}^T$

- **Z** (n x k)
  - principal component **scores**
- **V** (m x k)
  - **Loadings**
  - Principal component **coefficients**
  - Principal components

In PCA world, ***X*** is *n x m*

PCA minimizes Distortion

- **Mean center X, then compute the eigvectors**
  - $X^T X \, u_j$

PCA minimizes Distortion

- **First subtract off the average x from all the x_i**
  - From here, we'll assume this has been done
- **Approximate x in terms of an orthonormal basis v**
  - $\hat{x}_i = \Sigma_k \; z_{ik} \; \mathbf{v}_k$ or $\mathbf{X} = \mathbf{Z}\mathbf{V}^T$
- **Distortion** (this is the square of the earlier definition)

$$\sum_{i=1}^{n} ||\mathbf{x}^i - \hat{\mathbf{x}}^i||_2^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_j^i - \hat{x}_j^i)^2$$

PCA minimizes distortion

$$\mathbf{Distortion}_k : \sum_{i=1}^{n} \sum_{j=k+1}^{m} \mathbf{u}_j^\top (\mathbf{x}^i - \overline{\mathbf{x}})(\mathbf{x}^i - \overline{\mathbf{x}})^\top \mathbf{u}_j$$

$$= \sum_{j=k+1}^{m} \mathbf{u}_j^\top \left(\sum_{i=1}^{n} (\mathbf{x}^i - \overline{\mathbf{x}})(\mathbf{x}^i - \overline{\mathbf{x}})^\top \right) \mathbf{u}_j$$

$$= n \sum_{j=k+1}^{m} \mathbf{u}_j^\top \Sigma \mathbf{u}_j = n \sum_{j=k+1}^{m} \lambda_j$$

See the course wiki!

PCA maximizes variance

$$\textbf{Variance}_k : \sum_{i=1}^{n} \sum_{j=1}^{k} (\mathbf{u}_j^\top \mathbf{x}^i - \mathbf{u}_j^\top \overline{\mathbf{x}})^2$$

$$= \sum_{j=1}^{k} \mathbf{u}_j^\top \left(\sum_{i=1}^{n} (\mathbf{x}^i - \overline{\mathbf{x}})(\mathbf{x}^i - \overline{\mathbf{x}})^\top \right) \mathbf{u}_j$$

$$= n \sum_{j=1}^{k} \mathbf{u}_j^\top \Sigma \mathbf{u}_j .$$

See the course wiki!

PCA - Summary

$$\hat{\mathbf{x}}^i = \mathbf{x}^i = \overline{\mathbf{x}} + \sum_{j=1}^{m} z_j^i \mathbf{u}_j$$

$$\mathbf{Variance}_k + \mathbf{Distortion}_k = n \sum_{j=1}^{m} \lambda_j$$

See the course wiki!

Principal Component Analysis

$\mathbf{X} \longrightarrow \mathbf{X_c} = \mathbf{UDV^T} = \mathbf{ZV^T}$
$\mathbf{X_c}$ is (n x p), $\mathbf{Z}$ is (n x p), $\mathbf{V}$ is (p x p).

$\mathbf{Z}$ is the transformation of $\mathbf{X}$ into "PC space"
Column vector $\boldsymbol{z}_i$ is the i'th *PC score vector.*
Column vector $\boldsymbol{v}_i$ is the i'th *PC direction* or *loading.*

Since $\mathbf{V}$ is orthogonal, $\mathbf{X_c V} = \mathbf{ZV^TV} = \mathbf{Z}$, and therefore:

$$\boldsymbol{z}_i = \boldsymbol{X_c v}_i = \boldsymbol{u}_i D_{ii}$$

Hence $\boldsymbol{z}_i$ is the projection of the row vectors of $\mathbf{X_c}$ on the (unit) direction $\boldsymbol{v}_i$, scaled by D_{ii}.

Principal Component Analysis

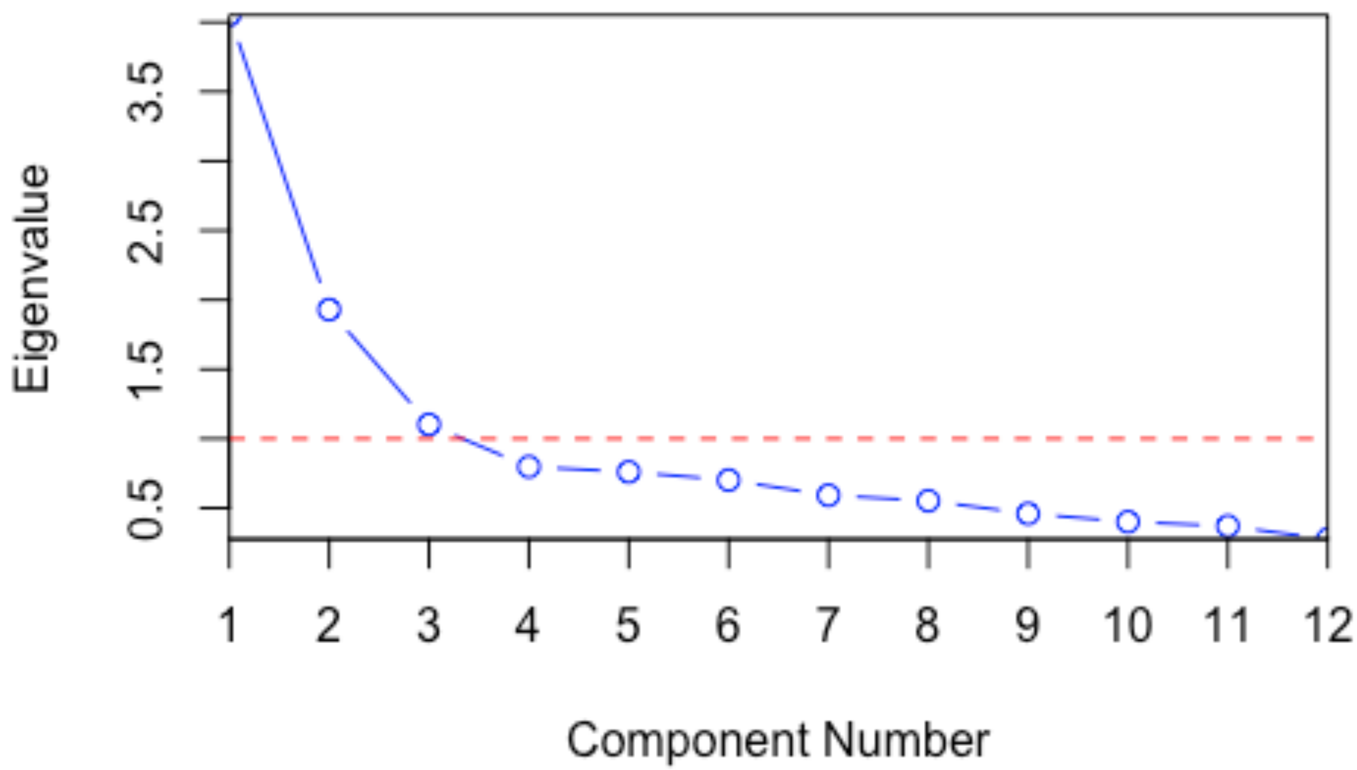
$$\mathrm{X} \longrightarrow \mathrm{X_c} = \mathrm{UDV^T} = \mathrm{ZV^T}$$

$$\boldsymbol{X_c^T X_c} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} (D_{ii})^2 \, \boldsymbol{v}_i \boldsymbol{v}_i^T$$

"% Variance explained by the i'th principal component:"

$$= 100 \cdot \frac{(D_{ii})^2}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} (D_{jj})^2} \qquad = 100\ \lambda_i / \Sigma_i\ \lambda_i$$

Scree plot



Keep components above the “elbow”

https://en.wikipedia.org/wiki/Scree_plot

PCA

True or false:

If **X** is any matrix, and **X** has singular value decomposition $\mathbf{X} = \mathbf{UDV}^T$ then the principal component scores for **X** are the columns of

$$\mathbf{Z} = \mathbf{UD}$$

a) True

b) False

PCA

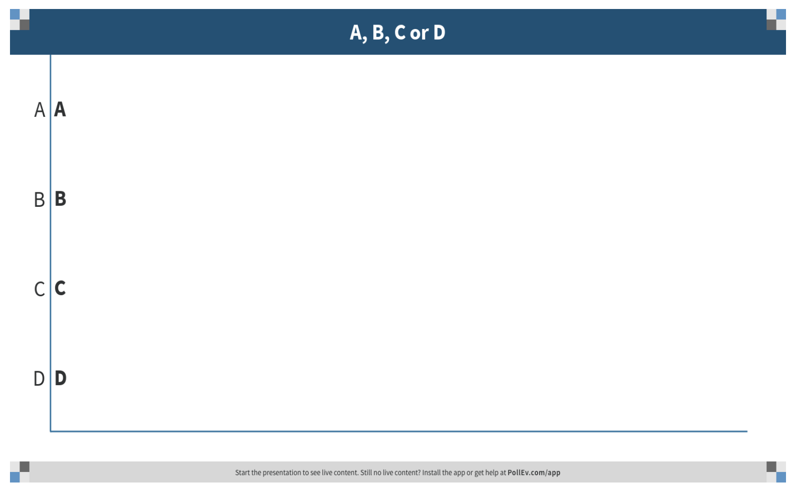
If X is mean-centered, then PCA finds…?

(a) Eigenvectors of $\mathbf{X}^T\mathbf{X}$

(b) Right singular vectors of $\mathbf{X}$

(c) Projection directions of maximum covariance of $\mathbf{X}$

(d) All of the above



PCA: Reconstruction Problem

PCA can be viewed as an L_2 optimization, minimizing distortion, the reconstruction error.

$$Z^*, V^* = \underset{\substack{Z \in \mathbb{R}^{n \times k},\, V \in \mathbb{R}^{p \times k}, \\ v_i^T v_j = \delta_{ij}}}{argmin} |X_c - ZV^T|_F$$

Here we have constrained **Z, V** by dimension:

$\mathbf{X_c}$ is still (n x p).

Z is (n x k), with k≤p.

V is (p x k).

If k=p then the reconstruction is perfect. k<p, not.

PCA via SVD

- **$X = ZV^T = UDV^T$**
  - **X** n x p **U** n x k **D** k x k **V^T** k x p
- **Z = UD - component scores or "factor scores”**
  - the transformed variable values corresponding to a particular data point
- **V^T - loadings**
  - the weight by which each standardized original variable should be multiplied to get the component score

PCA via SVD

- $\mathbf{x}_i = \Sigma_k z_{ik} \mathbf{v}_k$
- **What is z_{ik} ?**
  - $x_i = \Sigma_k u_{ik} d_{kk} v_k$

Sparse PCA

- $\text{argmin}_{\mathbf{Z},\mathbf{V}} ||\mathbf{X} - \mathbf{Z}\,\mathbf{V}^T||_F$
  - $\mathbf{v}_i'\mathbf{v}_j = \delta_{ij}$ (orthonormality)
  - Eigenvectors give the optimal solution
- **you can add an L_1 penalty**
- $\text{argmin}_{\mathbf{Z},\mathbf{V}} ||\mathbf{X} - \mathbf{Z}\,\mathbf{V}^T||_F + \lambda_1 ||\mathbf{Z}||_1 + \lambda_2 ||\mathbf{V}||_1$
  - No longer eigenvectors
  - Convex in **Z** given **V** and in **V** given **Z**
  - Solve by alternating gradient descent

What you should know

- **PCA as minimum reconstruction error (‘distortion’)**
- **PCA as finding direction of maximum covariance**
- **Sensitivity of PCA to standardizing**
- **Nomenclature:** scores, coefficients/loadings
- **Coming next: autoencoders, eigenfaces, eigenwords**