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, eigengrasp

PCA

- **PCA can be viewed as**
  - minimizing distortion $||\mathbf{x}_i - \Sigma_k z_{ik}\mathbf{v}_k||_2$
  - A rotation to a new coordinate system to maximize the variance in the new coordinates
- **Generally done by mean centering first**
  - You may or may not want to standardize

Nomenclature

X = ZV'

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

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**

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

PCA minimizes distortion

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

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

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

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

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

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

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, $\mathbf{X_cV} = \mathbf{ZV^TV} = \mathbf{Z}$, and therefore:

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

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

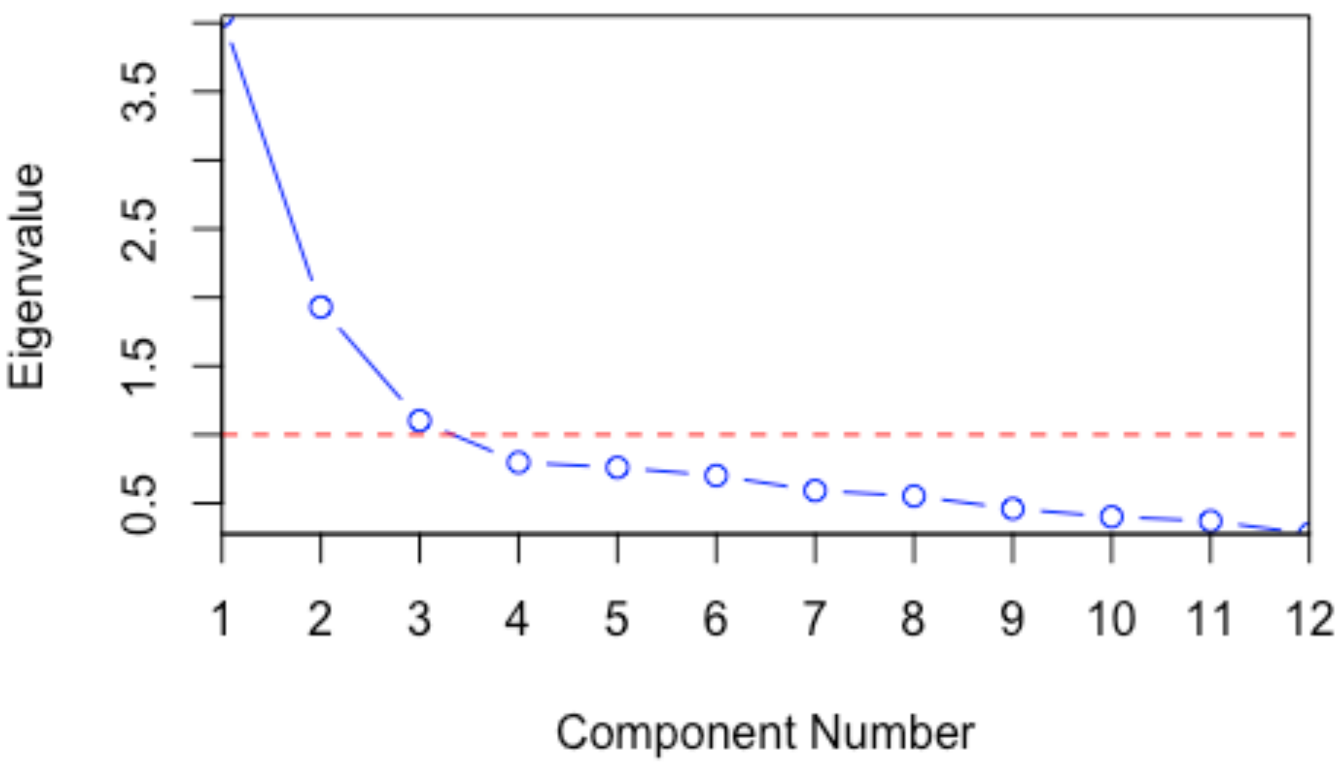
$$\mathbf{X} \longrightarrow \mathbf{X_c} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathbf{T}} = \mathbf{Z}\mathbf{V}^{\mathbf{T}}$$

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

"% 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 about the "elbow"

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

PCA

True or false:

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

$$Z = UD$$

a) True

b) False

PCA

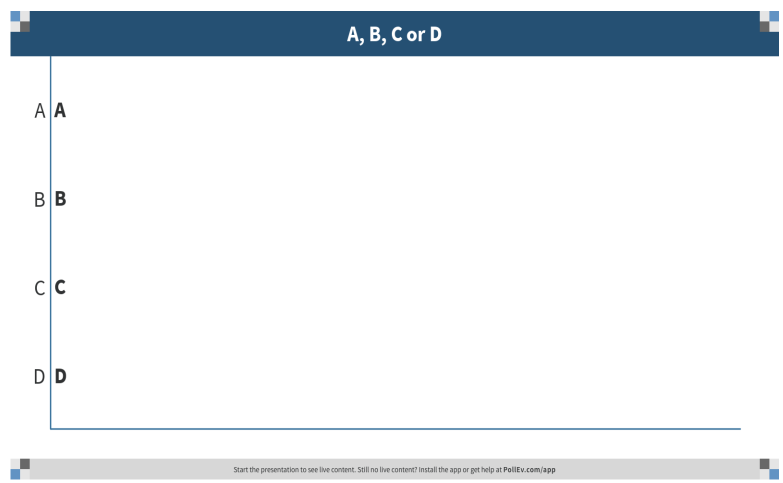
If X is mean-centered, then PCA finds…?

(a) Eigenvectors of X^TX

(b) Right singular vectors of X

(c) Projection directions of maximum covariance of 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

- **$X = ZV' = UDV'$**
  - **X** n x p **U** n x k **D** k x k **V'** k x p
- **$Z = UD$ - component scores or "factor scores"**
  - the transformed variable values corresponding to a particular data point
- **V' - 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

- $\text{argmin}_{Z,V} ||X - Z V'||_2$
  - $\mathbf{v}_i' \mathbf{v}_j = \delta_{ij}$ (orthonormality)
- with constraints
  - $||\mathbf{v}_i||_1 < c_1$ for i = 1...k
  - $||\mathbf{z}_i||_1 < c_2$ for i = 1...k
- or you can view this as a penalized regression – using Lagrange multipliers
- or you can use an L_1 penalty

PCR: Principal Component Regression

PCR has two steps:

1. Do a PCA on X to get component scores Z
2. Do OLS regression using Z as features

   $y = w'z$

PCR

- How to find z for a new x?
  - $X = ZV'$
- $xV = z V'V = z$

V	$p \times k$
$V'V=I$	$k \times k$
x	$1 \times p$
z	$1 \times k$

PCR: Principal Component Regression

$$\mathbf{X} \longrightarrow \mathbf{X_c} = \mathbf{ZV^T}$$

The columns $\mathbf{z_1},\ldots\mathbf{z_k}$ can be used as features in supervised learning.

Ex: linear regression. Given training **X** and **Y**,

$$w^* = \underset{w \in \mathbb{R}^p}{argmin} |Y - Zw|_2{}^2$$

If k=p: result is the *same* as linear regression with X**,** Y

If k<p: this is a form of *regularized* linear regression

So is ridge regression! How are PCR and Ridge fundamentally different?

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**