Updates

- HW 5 published
- Midterm regrades: later this week

Unsupervised Learning

- **Spectral methods**
  - Eigenvector/singular vector decomposition (SVD)
  - PCA, CCA
- **Reconstruction methods**
  - PCA, ICA, auto-encoders
- **Clustering and Probabilistic methods**
  - K-means
  - Gaussian mixtures
  - Latent Dirichlet Allocation (LDA)

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)**

$$x_i = \sum_k z_{ik} v_k$$

  - 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

- **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**
  - $\hat{x}_i = \Sigma_k \; z_{ik} \; v_k$ or $X = ZV^T$
- **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

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

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

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

See the course wiki!

PCA maximizes variance

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

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

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

See the course wiki!

PCA - Summary

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

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

See the course wiki!

Principal Component Analysis

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

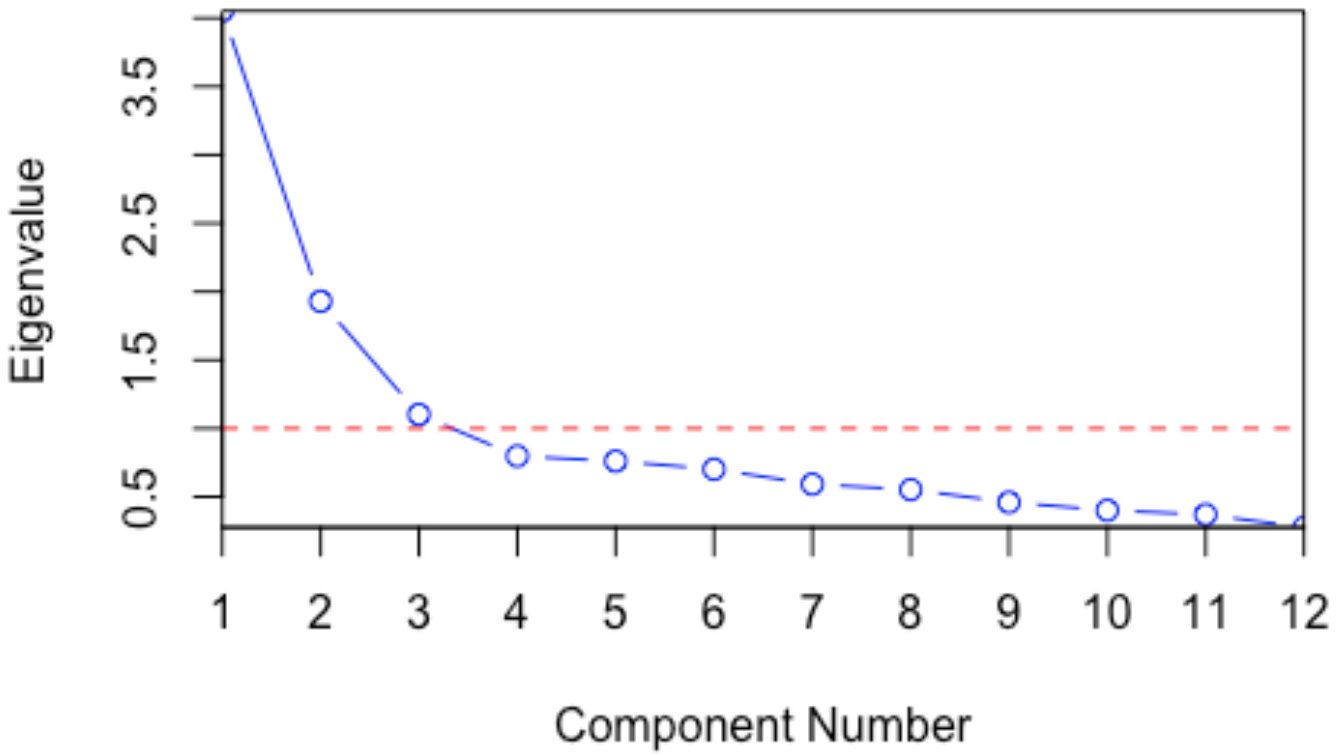
$$X \longrightarrow X_c = UDV^T = ZV^T$$

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

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

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

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 $X = UDV^T$

then the principal component scores for X are the columns of

$$Z = UD$$

a) True

b) False

PCA

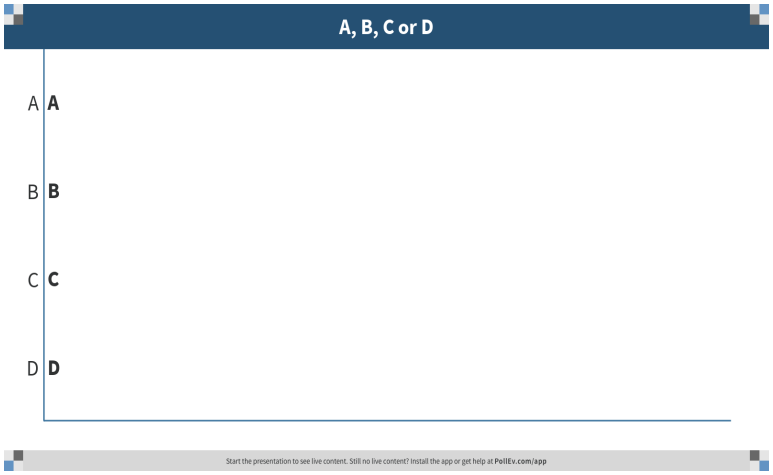
If X is mean-centered, then PCA finds…?

(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.

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

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

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

Z is (n x k), with $k \leq p$.

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**