Data Mining
S4240 Section 001

Giovanni Motta

Columbia University, Department of Statistics

September 23, 2015

Outline

Today:

1. High dimensional data
2. Principal components analysis (PCA) overview
3. (Review: eigenvalues and eigenvectors)
4. PCA computation

Next time: more PCA, doing PCA with R

High-Dimensional Data

Problem: want to describe student performance and compare students

Data ($n = 175$ students, $p = 8$ scores):

- scores on six homeworks
- midterm score
- final score

How can I summarize data for a student in a way that will be useful for comparisons?

Higher Dimensional Data

If data *truly* high dimensional:

- scores for each of 7 items have low correlation $\leftrightarrow$ need all items to summarize student
- example: good on HWs 1,2,3, middling on midterm, poor on HW 4, 5, and 6, great on final

But: we would not expect to see a student like the one above. Why?

Any suggestions for summarizing student performance in a way that is useful for comparing students? Why would this be reasonable?

Dimensionality Reduction

Data with n observations and p dimensions:

$$\underset{n\times p}{\mathbf{X}} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_p]$$

Here $\mathbf{x}^\top$ means the transpose of $\mathbf{x}$.

Today (and often, but not always), **bold** means matrix or vector, and *plain* means scalar.

Note that $\mathbf{X} \in \mathbb{R}^{n\times p}$

Dimensionality Reduction : $\kappa < p$

Two approaches:

- **Feature Selection:** choose a subset of features for prediction

$$\underbrace{[\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p]}_{p \text{ variables}} \rightarrow \underbrace{[\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_\kappa]}_{\kappa \text{ variables}}$$

  But you need to know your prediction task before you do feature selectionn.

- **Feature Extraction:** create new features by combining existing ones

$$\begin{aligned}[\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p] &\rightarrow [f_1(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p), \ldots, f_\kappa(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p)] \\ &= [\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_\kappa]\end{aligned}$$

Feature extraction is often a data preprocessing step since you do not need to know which prediction methods you will use.

Dimensionality Reduction

Today, we will look at *linear feature extraction*

$$\underset{p\times n}{\begin{bmatrix} \boldsymbol{X}_1^\top \\ \boldsymbol{X}_2^\top \\ \vdots \\ \boldsymbol{X}_j^\top \\ \vdots \\ \boldsymbol{X}_p^\top \end{bmatrix}} \rightarrow \underset{\kappa\times n}{\begin{bmatrix} \boldsymbol{Y}_1^\top \\ \boldsymbol{Y}_2^\top \\ \vdots \\ \boldsymbol{Y}_\kappa^\top \end{bmatrix}} = \underset{\kappa\times p}{\begin{bmatrix} w_{11} & w_{12} & \dots & w_{1j} & \dots & w_{1p} \\ w_{21} & w_{22} & \dots & w_{2j} & \dots & w_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ w_{\kappa 1} & w_{\kappa 2} & \dots & w_{\kappa j} & \dots & w_{\kappa p} \end{bmatrix}} \underset{p\times n}{\begin{bmatrix} \boldsymbol{X}_1^\top \\ \boldsymbol{X}_2^\top \\ \vdots \\ \boldsymbol{X}_j^\top \\ \vdots \\ \boldsymbol{X}_p^\top \end{bmatrix}}$$

$$\underset{\kappa\times n}{\mathbf{Y}^\top} = \underset{\kappa\times p}{\mathbf{W}^\top}\ \underset{p\times n}{\mathbf{X}^\top} \qquad \text{or} \qquad \underset{n\times\kappa}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\ \underset{p\times\kappa}{\mathbf{W}}$$

Terminology:

- $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_\kappa$ are the *scores*,
- $\mathbf{w}_1, \dots, \mathbf{w}_\kappa$ are the *loadings*.

Linear Feature Extraction

Student score averages are a form of linear feature extraction:

$$y_{i1} = [0.05,\, 0.05,\, 0.05,\, 0.05,\, 0.05,\, 0.05,\, 0.3,\, 0.4] \begin{bmatrix} HW1 \\ HW2 \\ HW3 \\ HW4 \\ HW5 \\ HW6 \\ Midterm \\ Final \end{bmatrix}$$

The student score average is the *score* and the assignment weights are the *loadings* (here $p = 8$ and $\kappa = 1$).

Linear Feature Extraction

How can we find a good set of loadings and scores for a general data set?

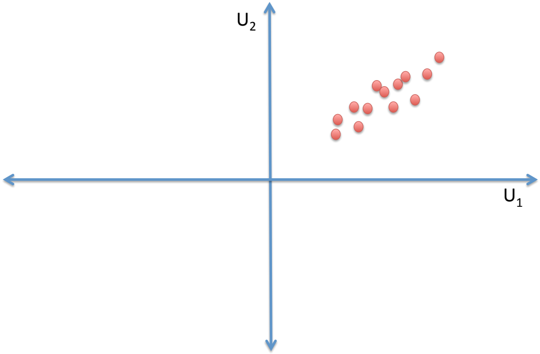
Principal components analysis (PCA): a covariance matrix singular value decomposition method for unsupervised data

Principal components: a set of linearly uncorrelated variables—these will be the loadings

Multiply loadings with original data to get scores!

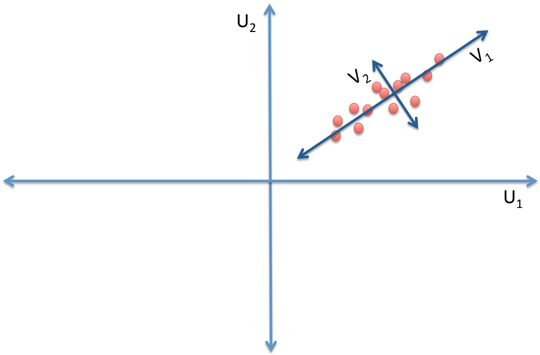
Principal Components Analysis

Basic idea: consider a dataset



Principal Components Analysis

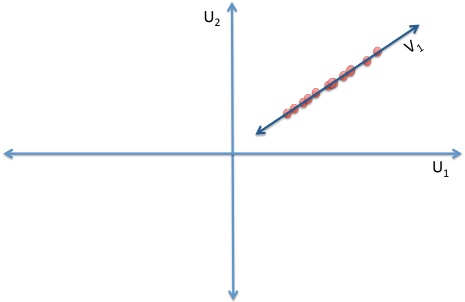
Basic idea: find a rotation that best describes data



A coordinate (linear) transformation from U to V!

Principal Components Analysis

Basic idea: given linear transformation, we can throw out the less descriptive dimensions and still have a decent representation



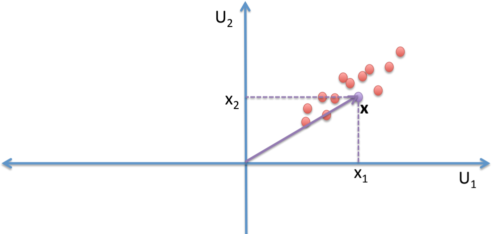
Consider the tree heights and weights. What is an interesting dimension? What is less descriptive?

Principal Components Analysis

Mathematically, how do we do this?

Simple case: $\mathbf{x} \in \mathbb{R}^2$ (i.e., $p = 2$), and we want a good projection $y \in \mathbb{R}$ (i.e., $\kappa = 1$)

- $\mathbf{x} = [x_1, x_2]^\top$, where x_1, x_2 are scalars
- the axes vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ are orthonormal
- $\mathbf{x} = x_1\mathbf{u}_1 + x_2\mathbf{u}_2 = \mathbf{I}_2\,\mathbf{x}$
- here $\mathbf{x}$ is one among the $p-$dimensional vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$



Principal Components Analysis

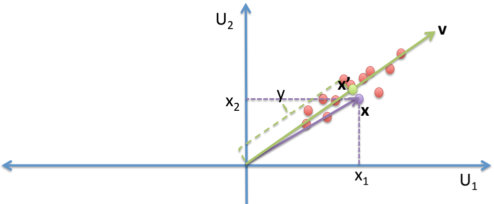
Mathematically, how do we do this?

Simple case: $\mathbf{x} \in \mathbb{R}^2$, and we want a good projection $y \in \mathbb{R}$

- we want to find a new direction vector $\mathbf{v}$
- let $\mathbf{w}_1$ be the linear transformation of $\mathbf{u}_1,\ \mathbf{u}_2$ into $\mathbf{v}$:

$$\mathbf{v} = w_{11}\mathbf{u}_1 + w_{12}\mathbf{u}_2 \ \ (= \mathbf{I}_2\,\mathbf{w}_1 = \mathbf{w}_1)$$

- Remember: $y_{ik} = \mathbf{x}_i^\top \mathbf{w}_k$, with $1 \le i \le n$ and $1 \le k \le \kappa$ Here $\kappa = 1$; $y := y_{i1} = \mathbf{x}^\top \mathbf{v} = \mathbf{x}^\top \mathbf{w}_1$
- project $\mathbf{x}$ onto $\mathbf{v}$ to make $\mathbf{x}' = y\,\mathbf{w}_1$ (y is the length of $\mathbf{x}'$)
- two coordinates x_1, x_2 become y (what have we lost?)



Principal Components Analysis

To find best $\mathbf{w}_1$:

- "closeness" is based on squared error between original points and new points
- usual notion of distance is Euclidean:

$$d(\mathbf{x}_i, \mathbf{x}'_i) = \sqrt{\sum_{j=1}^{p} \left(x_{ij} - x'_{ij}\right)^2}$$

- we will measure distance as Euclidean distance *squared* $\rightarrow$ big deviations heavily penalized, smaller deviations less so
- objective is to minimize squared errors

$$\hat{\mathbf{w}}_1 = \arg\min_{\mathbf{w}_1} \left\{ \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{x}'_i||_2^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} \left(x_{ij} - x'_{ij}\right)^2 \right\}$$

To find best $\mathbf{w}_1$:

- also have constraint:

$$\|\mathbf{w}_1\|_2 = \sqrt{\sum_{j=1}^{p} w_{ij}^2} = 1$$

this means one step in old coordinates $\mathbf{u}$ is equal to one step in new coordinates $\mathbf{v}$ (this is called the L2 norm... we will use this and the L1 norm $\|\mathbf{x}\|_1 = \sum_{j=1}^{p} |x_j|$ a lot in this class)

Principal Components Analysis

To find the best single feature $\mathbf{w}_1$:

1. center the data (subtract mean)... in R?
2. find best single feature, $\mathbf{w}_1$ by

$$
\begin{aligned}
\hat{\mathbf{w}}_1 &= \arg \min_{\mathbf{w}_1 \,:\, ||\mathbf{w}_1||_2 = 1} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(x_{ij} - x'_{ij}\right)^2 \\
&= \arg \min_{\mathbf{w}_1 \,:\, \|\mathbf{w}_1\|_2 = 1} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(x_{ij} - y_{i1} w_{1j}\right)^2 \\
&= \arg \min_{\mathbf{w}_1 \,:\, ||\mathbf{w}_1||_2 = 1} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(x_{ij} - w_{1j} \mathbf{x}_i^\top \mathbf{w}_1\right)^2 \\
&= \ldots \qquad \text{(using some matrix algebra } - \text{ see next slide)} \\
&= \arg \max_{\mathbf{w}_1 \,:\, ||\mathbf{w}_1||_2 = 1} \left\{ \mathbf{w}_1^\top \mathbf{X}^\top \mathbf{X} \mathbf{w}_1 = \frac{\mathbf{w}_1^\top \left(\mathbf{X}^\top \mathbf{X}\right) \mathbf{w}_1}{\mathbf{w}_1^\top \mathbf{w}_1} \right\}
\end{aligned}
$$

Principal Components Analysis

$$
\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{ij}-x'_{ij}\right)^2 &= \sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{ij}-w_{1j}\mathbf{x}_i^\top\mathbf{w}_1\right)^2 \\
\left\|\mathbf{X}-\mathbf{X}\mathbf{w}_1\mathbf{w}_1^\top\right\|_2^2 &= \left\|\mathbf{X}(\mathbf{I}_p-\mathbf{w}_1\mathbf{w}_1^\top)\right\|_2^2 \\
\mathrm{tr}[(\mathbf{I}_\mathrm{p}-\mathbf{w}_1\mathbf{w}_1^\top)^\top\mathbf{X}^\top\mathbf{X}(\mathbf{I}_\mathrm{p}-\mathbf{w}_1\mathbf{w}_1^\top)] &= \mathrm{tr}[(\mathbf{X}^\top\mathbf{X}-\mathbf{w}_1\mathbf{w}_1^\top\mathbf{X}^\top\mathbf{X})(\mathbf{I}_\mathrm{p}-\mathbf{w}_1\mathbf{w}_1^\top)]
\end{aligned}
$$

$$
= \mathrm{tr}[\mathbf{X}^\top\mathbf{X}]-\mathrm{tr}[\mathbf{X}^\top\mathbf{X}\mathbf{w}_1\mathbf{w}_1^\top]-\mathrm{tr}[\mathbf{w}_1\mathbf{w}_1^\top\mathbf{X}^\top\mathbf{X}]+\mathrm{tr}[\mathbf{w}_1\mathbf{w}_1^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}_1\mathbf{w}_1^\top]
$$

(using the cyclical property of the trace and $\mathbf{w}_1^\top\mathbf{w}_1 = 1$)

$$
\begin{aligned}
&= \mathrm{tr}[\mathbf{X}^\top\mathbf{X}]-2\mathrm{tr}[\mathbf{w}_1^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}_1]+\mathrm{tr}[\mathbf{w}_1^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}_1] \\
&= \mathrm{tr}[\mathbf{X}^\top\mathbf{X}]-\mathrm{tr}[\mathbf{w}_1^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}_1] \\
&= \mathrm{tr}[\mathbf{X}^\top\mathbf{X}]-\mathbf{w}_1^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}_1
\end{aligned}
$$

Principal Components Analysis

An important fact: $\frac{1}{n}\mathbf{X}^\top\mathbf{X} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$ is the maximum likelihood estimate of the covariance matrix for $[x_1, \ldots, x_p]^\top$

The scalar

$$\frac{\mathbf{w}_1^\top \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\mathbf{w}_1}{\mathbf{w}_1^\top\mathbf{w}_1} \;\left(= \frac{1}{n}\boldsymbol{Y}_1^\top\boldsymbol{Y}_1 = \frac{1}{n}\sum_{i=1}^n y_{i1}^2 \right)$$

is called a *Rayleigh quotient*

Another fact: since $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$ is symmetric,

- the value of $\mathbf{w}_1$ that maximizes the Rayleigh quotient is the eigenvector associated with the largest eigenvalue of $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$
- the corresponding maximum of the Rayleigh quotient is the largest eigenvalue of $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$

Principal Components Analysis

- center the data (column-wise) and redefine $\mathbf{X} = \mathbf{H}\mathbf{X}$, where $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$
- Let $\boldsymbol{Y}_1 = \mathbf{X}\mathbf{w}_1$, and rewrite the problem as

$$\max_{\mathbf{w}_1^\top \mathbf{w}_1 = 1} \|\boldsymbol{Y}_1\| = \max_{\mathbf{w}_1^\top \mathbf{w}_1 = 1} \mathbf{w}_1^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}_1$$

- Lagrange function $L_1(\mathbf{w}_1, \mathbf{X}, \lambda) = \mathbf{w}_1^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}_1 - \lambda(\mathbf{w}_1^\top \mathbf{w}_1 - 1)$
- Our problem is equivalent to

$$\max_{\mathbf{w}_1} L_1(\mathbf{w}_1, \mathbf{X}, \lambda)$$

- solution

$$\frac{\partial L_1}{\partial \mathbf{w}_1} = 2\mathbf{X}^\top \mathbf{X}\mathbf{w}_1 - 2\lambda\mathbf{w}_1 = 0 \Longleftrightarrow \mathbf{X}^\top \mathbf{X}\mathbf{w}_1 = \lambda_1 \mathbf{w}_1$$

How to find the second new direction?

Principal Components Analysis

- Let $\boldsymbol{Y}_2 = \mathbf{X}\mathbf{w}_2$, and rewrite the problem as

$$\max_{\mathbf{w}_2^\top \mathbf{w}_2 = 1,\ \mathbf{w}_2^\top \mathbf{w}_1 = 0} \|\boldsymbol{Y}_2\| = \max_{\mathbf{w}_2^\top \mathbf{w}_2 = 1,\ \mathbf{w}_2^\top \mathbf{w}_1 = 0} \mathbf{w}_2^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}_2$$

- Lagrange function
  $L_2(\mathbf{w}_2, \mathbf{w}_1, \mathbf{X}, \lambda) = \mathbf{w}_2^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}_2 - \lambda(\mathbf{w}_2^\top \mathbf{w}_2 - 1) - \mu(\mathbf{w}_2^\top \mathbf{w}_1)$
- Our problem is equivalent to

$$\max_{\mathbf{w}_2} L_2(\mathbf{w}_2, \mathbf{w}_1, \mathbf{X}, \lambda)$$

- solution

$$\frac{\partial L_2}{\partial \mathbf{w}_2} = 2\mathbf{X}^\top \mathbf{X}\mathbf{w}_2 - 2\lambda \mathbf{w}_2 - \mu \mathbf{w}_1 = 0$$

  Pre-multiply by $\mathbf{w}_2^\top$

  $2\mathbf{w}_2^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}_2 - 2\lambda \mathbf{w}_2^\top \mathbf{w}_2 - \mu \mathbf{w}_2^\top \mathbf{w}_1 = 0 \Longleftrightarrow \mathbf{X}^\top \mathbf{X}\mathbf{w}_2 = \lambda_2 \mathbf{w}_2$

How to find the third new direction? Solution: $\mathbf{X}^\top \mathbf{X}\mathbf{w}_3 = \lambda_3 \mathbf{w}_3$
... and so on ... up to the p-th eigenvector (the one corresponding to the smallest eigenvalues)

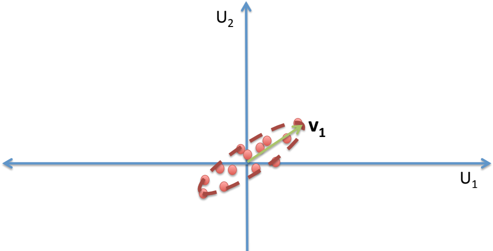
Principal Components Analysis

Geometric interpretation.

- objective is to minimize squared errors

$$\hat{\mathbf{w}}_1 = \arg \min_{\mathbf{w}_1 : ||\mathbf{w}_1||_2 = 1} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(x_{ij} - x'_{ij}\right)_2^2$$

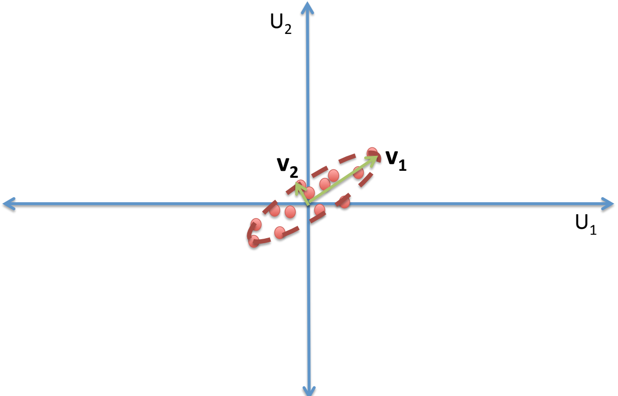
- center data
- fit a multivariate Gaussian distribution fit to the data
- Gaussian has mean $\mathbf{0}$, covariance $\mathbf{\Sigma}$
- $\mathbf{w}_1$ is the direction of the covariance ellipse with maximum variance



Principal Components Analysis

This generalizes to multiple dimensions

- suppose we have p original dimensions and we would like κ new ones
- the optimal vectors have the same direction as the κ ellipse directions with maximum variance



Principal Components Analysis

How do we find the κ most descriptive directions?

- center the p variables to that they all have mean 0
- estimate the covariance matrix

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top$$

- find the *eigenvalues* and *eigenvectors* of the covariance matrix
- the κ most descriptive directions are the eigenvectors associated with the κ largest eigenvalues

The Truth: Matrices, Eigenvalues, and Eigenvectors

Matrices can be used to describe transformations:

- under transformations, some of the original directions may be preserved

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

- the direction $\mathbf{x}$ is an *eigenvector* of $\mathbf{A}$
- the scalar λ is an *eigenvalue* $\mathbf{A}$
- the eigenvalue describes the magnitude of the change in $\mathbf{x}$ under $\mathbf{A}$

Let's play with an animated gif [eigshow in `MATLAB`]. `What is the matrix? What are the eigenvectors? What are the eigenvalues?`

The Truth: Matrices, Eigenvalues, and Eigenvectors

Matrices can be used to describe transformations:

- $\mathbf{x} \rightarrow \mathbf{A}\mathbf{x}$ is a transformation of $\mathbf{x}$
- Examples (look at these graphically):
  - $\mathbf{x} =$ $(1,1)$, $(1,0)$, $(1,-1)$, $(0,1)$, $(-1,1)$, $(-1,0)$ $(-1,1)$, $(0,-1)$
  - $\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$
  - $\mathbf{A} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

The Truth: Matrices, Eigenvalues, and Eigenvectors

To find the eigenvectors of $\mathbf{A}$, find the roots of the polynomial

$$\det\left(\mathbf{A} - \lambda I\right) = 0$$

Let's do this with our examples:

- $\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

- $\mathbf{A} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

The Truth: Matrices, Eigenvalues, and Eigenvectors

We can use the eigenvalues to find the eigenvectors

- start with eigenvalue λ_i
- solve the set of linear equations

$$\mathbf{A}\mathbf{x} = \lambda_i \mathbf{x}$$

- $\mathbf{x}$ is the eigenvector associated with λ_i

Let's do this with our examples:

- $\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

- $\mathbf{A} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

The Truth: Matrices, Eigenvalues, and Eigenvectors

We have 2 matrices:

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.$$

Let's find the eigenvalues:

$$\begin{aligned} 0 &= \det(\mathbf{A} - \lambda I) \\ &= \det\left(\begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix}\right) \\ &= (2-\lambda)^2 - 1 \\ &= (\lambda - 3)(\lambda - 1) \end{aligned}$$

$$\begin{aligned} 0 &= \det(\mathbf{B} - \lambda I) \\ &= \det\left(\begin{bmatrix} 2-\lambda & 0 \\ 0 & 2-\lambda \end{bmatrix}\right) \\ &= (2-\lambda)^2 \end{aligned}$$

So the eigenvalues are 3 and 1 for $\mathbf{A}$; 2 and 2 for $\mathbf{B}$.

The Truth: Matrices, Eigenvalues, and Eigenvectors

Now use the eigenvalues to find the eigenvectors: $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$.

Solve for $\mathbf{A}$:

$$\mathbf{A}\mathbf{x}_1 = 3\mathbf{x}_1$$

$$\begin{bmatrix} 2x_{11} + x_{12} \\ x_{11} + 2x_{12} \end{bmatrix} = \begin{bmatrix} 3x_{11} \\ 3x_{12} \end{bmatrix}$$

$$\mathbf{x}_1 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$$

$$\mathbf{A}\mathbf{x}_2 = 1\mathbf{x}_2$$

$$\begin{bmatrix} 2x_{21} + x_{22} \\ x_{21} + 2x_{22} \end{bmatrix} = \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix}$$

$$\mathbf{x}_2 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix}$$

Solve for $\mathbf{B}$:

$$\mathbf{B}\mathbf{x}_1 = 2\mathbf{x}_1$$

$$\begin{bmatrix} 2x_{11} \\ 2x_{12} \end{bmatrix} = \begin{bmatrix} 2x_{11} \\ 2x_{12} \end{bmatrix}$$

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ or } \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Principal Components Analysis

Back to PCA...

1. center the data
2. compute $\mathbf{X}^\top \mathbf{X}$
3. compute the κ eigenvectors corresponding to the largest κ eigenvalues of $\mathbf{X}^\top \mathbf{X}$ to get loadings
4. for the eigenvectors with the κ largest eigenvalues, make factor scores by setting

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{i\kappa} \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1p} \\ w_{21} & w_{22} & \dots & w_{2p} \\ \vdots & & & \vdots \\ w_{\kappa 1} & w_{\kappa 2} & \dots & w_{\kappa p} \end{bmatrix} \begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ \vdots \\ \vdots \\ x_{ip} \end{bmatrix}$$

Principal Components Analysis

5. do computations in the (smaller, κ-dimensional) space, the space of factor scores $\mathbf{y}$
6. transform results back to original space (if needed)

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{i\kappa} \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1p} \\ w_{21} & w_{22} & \dots & w_{2p} \\ \vdots & & & \vdots \\ w_{\kappa 1} & w_{\kappa 2} & \dots & w_{\kappa p} \end{bmatrix} \begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ \vdots \\ \vdots \\ x_{ip} \end{bmatrix}$$

Principal Components Analysis

Let's do an example:

$$\mathbf{X} = \begin{bmatrix} -4 & -4 \\ -1 & 1 \\ 1 & -1 \\ 4 & 4 \end{bmatrix}$$

$$[\mathbf{w}_1, \mathbf{w}_2] =$$

$$[\mathbf{y}_1, \mathbf{y}_2] =$$

Let's find the eigenvalues and eigenvectors of the empirical covariance matrix of $\mathbf{X}$ for PCA.

PCA: steps to find the optimal rotation

1. Is the data centered? Yes, the sum of all of the columns of $\mathbf{X}$ is 0.)
2. Find the empirical covariance matrix $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$:

$$\frac{1}{n}\mathbf{X}^\top\mathbf{X} = \frac{1}{4}\begin{bmatrix} 34 & 30 \\ 30 & 34 \end{bmatrix}$$

3. Now find the eigenvalues:

$$\begin{aligned} 0 &= \det\left(\begin{bmatrix} \frac{34}{4} - \lambda & \frac{30}{4} \\ \frac{30}{4} & \frac{34}{4} - \lambda \end{bmatrix}\right) \\ &= (\lambda - 16)(\lambda - 1) \end{aligned}$$

Therefore $\lambda = 16$ is the eigenvalue associated with the most expressive linear rotation, $\lambda = 1$ is the second the eigenvalue associated with the second most expressive.

PCA: steps to find the optimal rotation

4. Find the eigenvectors associated with those eigenvalues:

$$\begin{aligned} \tfrac{1}{4}\mathbf{X}^\top\mathbf{X}\mathbf{w}_1 &= 16\mathbf{w}_1 \\ \begin{bmatrix} \frac{34}{4}w_{11} + \frac{30}{4}w_{12} \\ \frac{30}{4}w_{11} + \frac{34}{4}w_{12} \end{bmatrix} &= \begin{bmatrix} 16w_{11} \\ 16w_{12} \end{bmatrix} \end{aligned}$$

$$\leftrightarrow \mathbf{w}_1 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$$

$$\tfrac{1}{4}\mathbf{X}^\top\mathbf{X}\mathbf{w}_2 = 1\mathbf{w}_2$$

$$\begin{bmatrix} \frac{34}{4}w_{21} + \frac{30}{4}w_{22} \\ \frac{30}{4}w_{21} + \frac{34}{4}w_{22} \end{bmatrix} = \begin{bmatrix} w_{21} \\ w_{22} \end{bmatrix}$$

$$\leftrightarrow \mathbf{w}_2 = \begin{bmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$$

Note that the eigenvectors have magnitude 1:

$$\sum_{j=1}^{p} w_{jk}^2 = 1, \qquad 1 \leq k \leq p.$$

This means that one step is the new directions is equivalent to one step in the old directions.

PCA: steps to find the optimal rotation

5. The eigenvectors make up your loading matrix, $\mathbf{W}$:

$$\mathbf{W} = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$$

   Since eigenvectors form an orthonormal basis and have magnitude 1, $\mathbf{W}^\top \mathbf{W} = \mathbf{I}$.

6. Find the scores for each data point by $\mathbf{Y} = \mathbf{X}\mathbf{W}$.