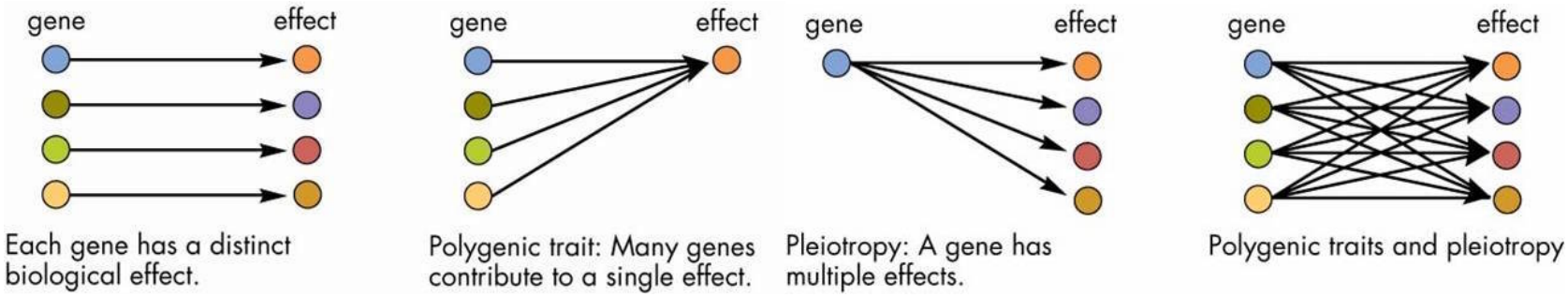
Latent Variable Models

Schwartz

August 13, 2017

Genetics Primer

- Polygenic Inheritance: trait is controlled by 2+ genes
- Pleiotropy: gene drives 2+ seemingly unrelated traits



$$\begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_2^2 \end{bmatrix}$$

Experimental Design

$$f(Y|\boldsymbol{X})$$

Predictive Modeling

$$\begin{bmatrix} 1^2 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1^2 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1^2 \end{bmatrix}$$

Correlated Features

PCA
FA
ICA

Objectives

- Covariance Matrices Vs. Correlation Matrices
  - And what they do
- Eigen-vectors/values of Correlation & Covariance Matrices
  - And what they do:
  - Proportion of Variance Explained
    - Skree Plot
  - Directions
  - Projecting onto eigenvectors
- Principal Components Analysis (PCA)
  - Interpretation
    - Biplots
  - Dimensionality Reduction
  - Visualization
- Singular Value Decomposition (SVD)
  - Same, but better than PCA?
- Factor Analysis (FA)
- Independent Component Analysis (ICA)

Sample Covariance Matrix

$$\boldsymbol{X}_{n \times p}:$$

Feature	X_1	X_2	$\cdots$	X_p
n=1				
2				
$\vdots$				
n				

Sample Covariance Matrix

$\boldsymbol{X}_{n \times p}$:

Feature	X_1	X_2	$\cdots$	X_p
n=1				
2				
$\vdots$				
n				

the sample *covariance matrix* and *correlation matrix* capture observed feature covariances and correlations, respectively

$\boldsymbol{\Sigma}_{\mathbf{p} \times \mathbf{p}}$	X_1	X_2	$\cdots$	X_p
X_1	σ^2_{11}	σ^2_{12}		σ^2_{1p}
X_2	σ^2_{21}	σ^2_{22}		σ^2_{2p}
$\vdots$		$\ddots$		
X_p	σ^2_{p1}	σ^2_{p2}		σ^2_{pp}

$\mathbf{r}_{\mathbf{p} \times \mathbf{p}}$	X_1	X_2	$\cdots$	X_p
X_1	1	r_{12}		r_{1p}
X_2	r_{21}	1		r_{2p}
$\vdots$		$\ddots$		
X_p	r_{p1}	r_{p2}		1

Sample Covariance Matrix

$\boldsymbol{X}_{n \times p}$:

Feature	X_1	X_2	$\cdots$	X_p
n=1				
2				
$\vdots$				
n				

the sample *covariance matrix* and *correlation matrix* capture observed feature covariances and correlations, respectively

$\boldsymbol{\Sigma}_{\mathbf{p} \times \mathbf{p}}$	X_1	X_2	$\cdots$	X_p
X_1	σ^2_{11}	σ^2_{12}		σ^2_{1p}
X_2	σ^2_{21}	σ^2_{22}		σ^2_{2p}
$\vdots$		$\ddots$		
X_p	σ^2_{p1}	σ^2_{p2}		σ^2_{pp}

$\mathbf{r}_{\mathbf{p} \times \mathbf{p}}$	X_1	X_2	$\cdots$	X_p
X_1	1	r_{12}		r_{1p}
X_2	r_{21}	1		r_{2p}
$\vdots$		$\ddots$		
X_p	r_{p1}	r_{p2}		1

For a **centered** samples by features data matrix ($X_{ij} = X^*_{ij} - \bar{X}^*_j$)

$$\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}^T\boldsymbol{X} = \frac{1}{n-1}\sum_{i=1}^{n} X_{i\cdot}X_{i\cdot}^T$$

$$\hat{\boldsymbol{r}} = \hat{\boldsymbol{D}}_S^{-\frac{1}{2}}\hat{\boldsymbol{S}}\hat{\boldsymbol{D}}_S^{-\frac{1}{2}}$$

$Cov(X_j, X_{j'}) = \hat{S}_{jj'} = \frac{1}{n-1}\sum_{i=1}^{n} X_{ij}X_{ij'}$

D_S is the $p \times p$ matrix with same diagonal as S

Sample Covariance Matrix

$\boldsymbol{X}_{n \times p}$:

Feature	X_1	X_2	$\cdots$	X_p
n=1				
2				
$\vdots$				
n				

the sample *covariance matrix* and *correlation matrix* capture observed feature covariances and correlations, respectively

$\mathbf{S_{p \times p}}$	X_1	X_2	$\cdots$	X_p
X_1	$\hat{\sigma}^2_{11}$	$\hat{\sigma}^2_{12}$		$\hat{\sigma}^2_{1p}$
X_2	$\hat{\sigma}^2_{21}$	$\hat{\sigma}^2_{22}$		$\hat{\sigma}^2_{2p}$
$\vdots$		$\ddots$		
X_p	$\hat{\sigma}^2_{p1}$	$\hat{\sigma}^2_{p2}$		$\hat{\sigma}^2_{pp}$

$\hat{\mathbf{r}}_{\mathbf{p \times p}}$	X_1	X_2	$\cdots$	X_p
X_1	1	$\hat{r}_{12}$		$\hat{r}_{1p}$
X_2	$\hat{r}_{21}$	1		$\hat{r}_{2p}$
$\vdots$		$\ddots$		
X_p	$\hat{r}_{p1}$	$\hat{r}_{p2}$		1

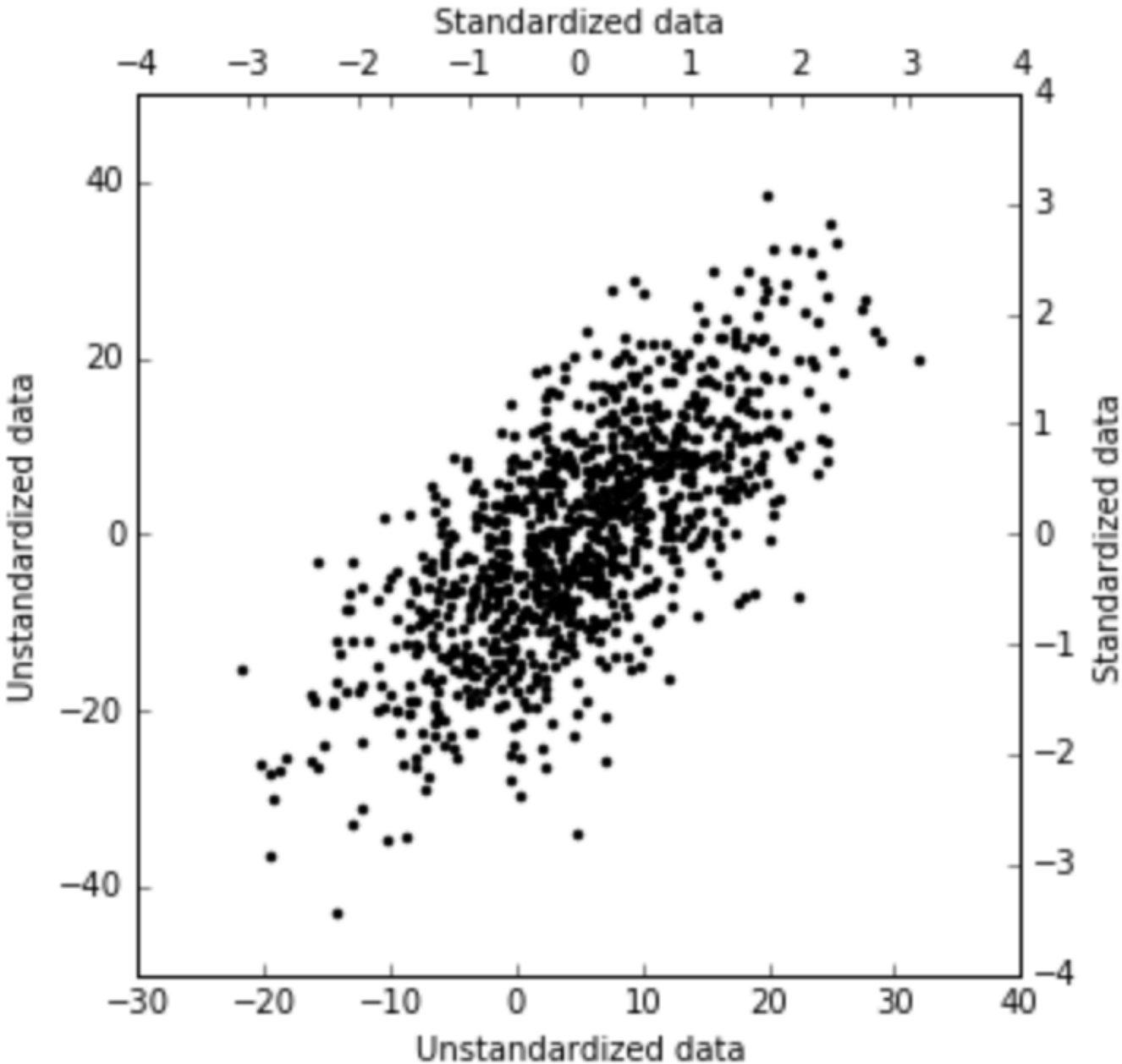
For a **centered** samples by features data matrix ($X_{ij} = X^*_{ij} - \bar{X}^*_j$)

$$\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}^T\boldsymbol{X} = \frac{1}{n-1}\sum_{i=1}^{n} X_{i\cdot}X_{i\cdot}^T \qquad \hat{\boldsymbol{r}} = \hat{\boldsymbol{D}}_S^{-\frac{1}{2}}\hat{\boldsymbol{S}}\hat{\boldsymbol{D}}_S^{-\frac{1}{2}}$$

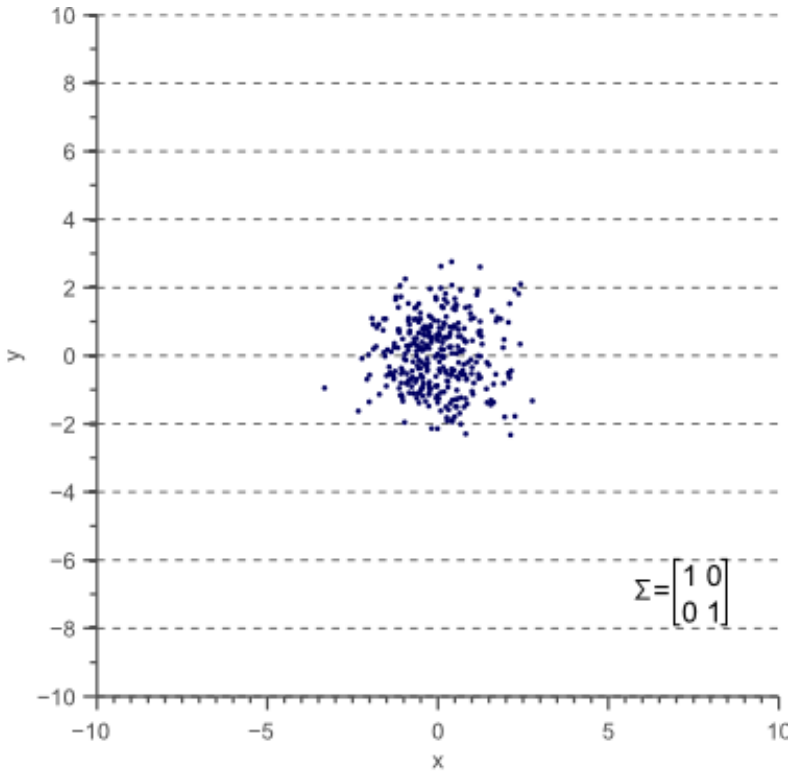
$Cov(X_j, X_{j'}) = \hat{S}_{jj'} = \frac{1}{n-1}\sum_{i=1}^{n} X_{ij}X_{ij'}$

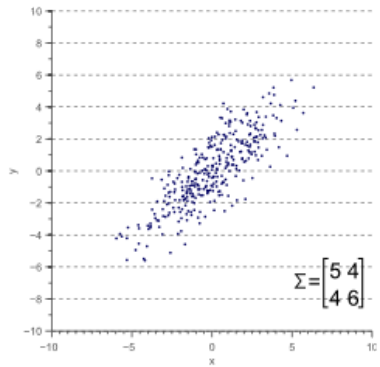
D_S is the $p \times p$ matrix with same diagonal as S

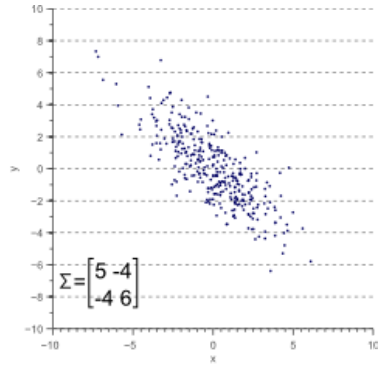
Covariance versus Correlation

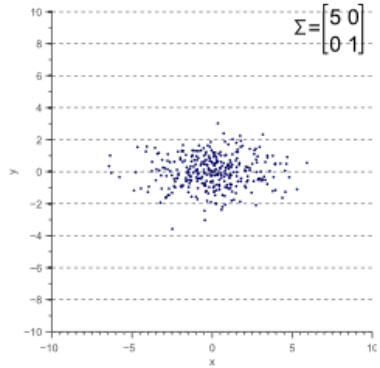


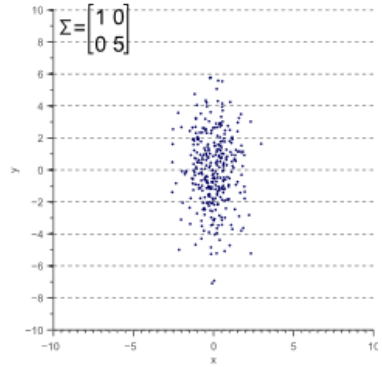
Sample Covariance (Correlation) Linear Transformation

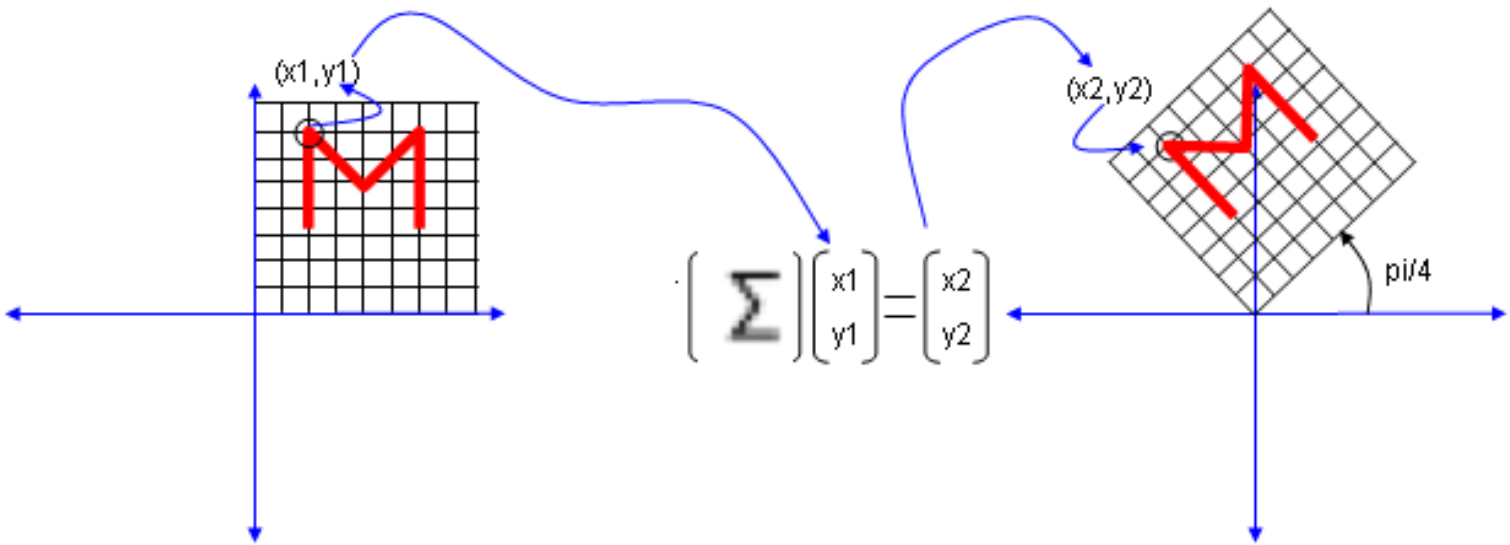












Eigenvectors and Eigenvalues

The *eigenvectors and eigenvalues* of a linear transformation $\mathbf{\Sigma}_{p \times p}$ are orthonormal vectors v and constants λ satisfying

$$\mathbf{\Sigma} v = \lambda v$$

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Eigenvectors and Eigenvalues

The *eigenvectors and eigenvalues* of a linear transformation $\mathbf{\Sigma}_{p\times p}$ are orthonormal vectors v and constants λ satisfying

$$\mathbf{\Sigma}v = \lambda v \quad \text{as opposed to} \quad \mathbf{\Sigma}w = w'$$

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

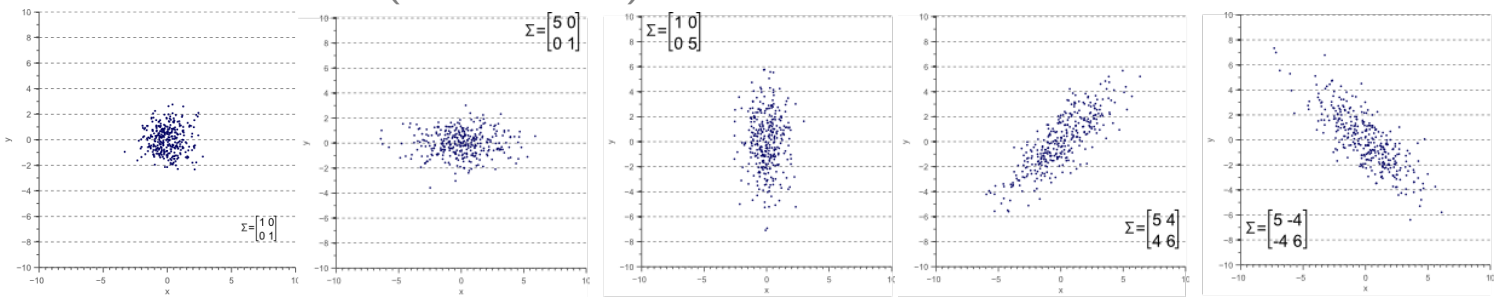
Eigenvectors and Eigenvalues

The *eigenvectors and eigenvalues* of a linear transformation $\mathbf{\Sigma}_{p\times p}$ are orthonormal vectors v and constants λ satisfying

$$\mathbf{\Sigma}v = \lambda v \quad \text{as opposed to} \quad \mathbf{\Sigma}w = w'$$

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

The covariance (correlation) matrix correlates uncorrelated variables



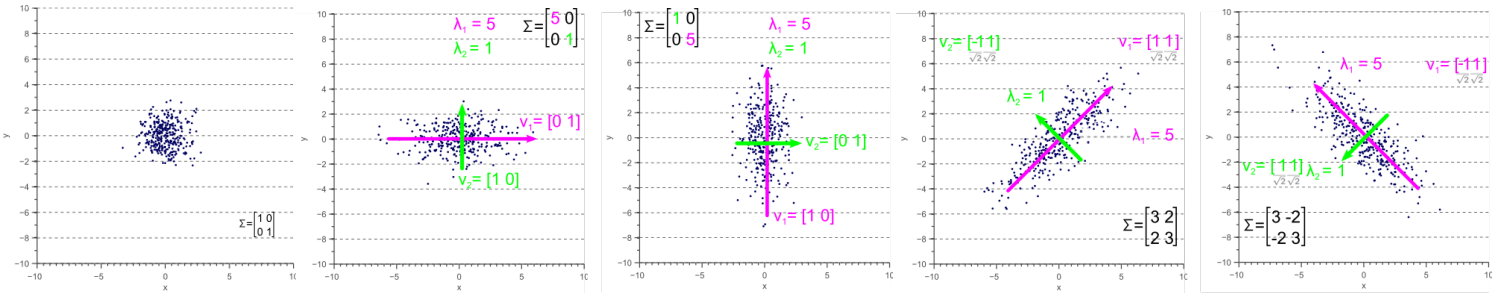
Eigenvectors and Eigenvalues

The *eigenvectors* and *eigenvalues* of a linear transformation $\mathbf{\Sigma}_{p\times p}$ are orthonormal vectors v and constants λ satisfying

$$\mathbf{\Sigma}v = \lambda v \quad \text{as opposed to} \quad \mathbf{\Sigma}w = w'$$

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = 2\begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

The covariance (correlation) matrix correlates uncorrelated variables



The *eigenvectors* of the covariance (correlation) matrix turn out to be unit vectors v_j sequentially maximizing the variance of *scores* $z_{ij} = v_j^T X_{i\cdot}$ that are sequentially uncorrelated with previous *scores*.

I.e. for each j, v_j solve $\underset{v_j}{argmax} \sum_{j'=1}^{j} Var(z_{ij'}) = \sum_{j'=1}^{j} \lambda_{j'}$ such that $Cov(z_{ij}, z_{ij'}) = 0$ for $j' < j$, with λ_j the eigenvalue of v_j.

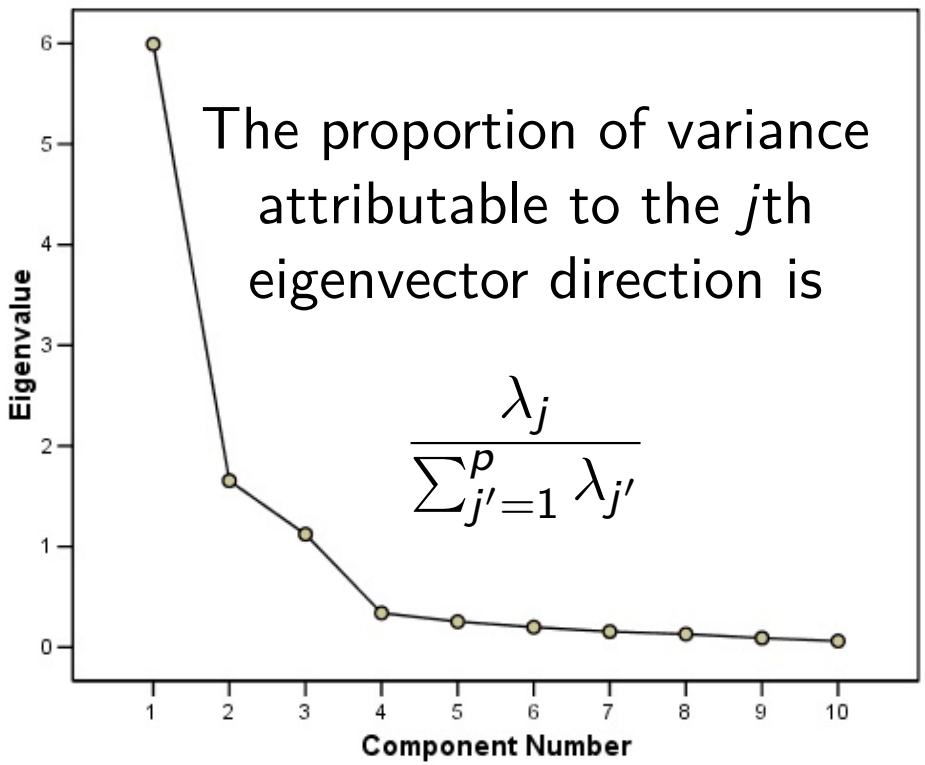
Proportion of Variation Explained

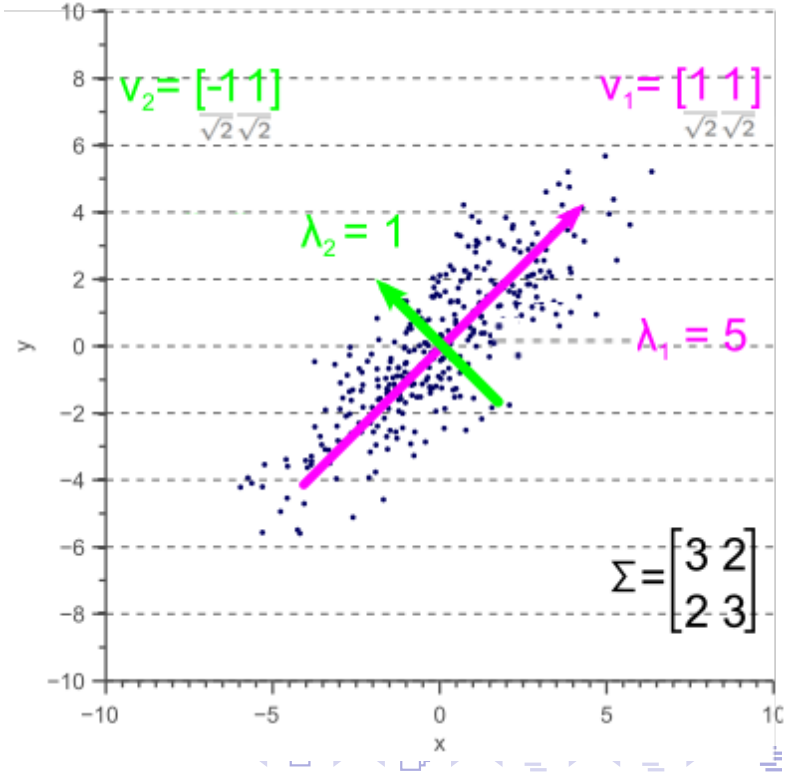
Since our eigenvalues are orthonormal, i.e.,

$$\sum_{k=1}^{p} v_{kj}^2 = 1 \text{ and } v_j^T v_{j'} = 0 \text{ for } j \neq j' \text{ we have that}$$

$$\sum_{j=1}^{p} Var_i(X_{ij}) = \sum_{j=1}^{p} Var_i(v_j^T X_{i\cdot}) = \sum_{j=1}^{p} Var_i(z_{ij}) = \sum_{j=1}^{p} \lambda_j$$

I.e., data variance cay be partitioned between eigenvector directions





Biasing the "Proportion of Variance Explained" *Direction*

The eigenvalues point towards the most correlated axes, e.g.

$$\begin{bmatrix} 1 & 0.9 & 0.1 & 0.1 \\ 0.9 & 1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1 & 0.9 \\ 0.1 & 0.1 & 0.9 & 1 \end{bmatrix} \quad v_1 = 1/\sqrt{\frac{1}{4}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad v_2 = 1/\sqrt{\frac{1}{4}} \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & .6 & .5 \\ .6 & 1 & .4 \\ .5 & .4 & 1 \end{bmatrix} \quad v_1 = \begin{bmatrix} 0.61 \\ 0.58 \\ 0.54 \end{bmatrix}$$

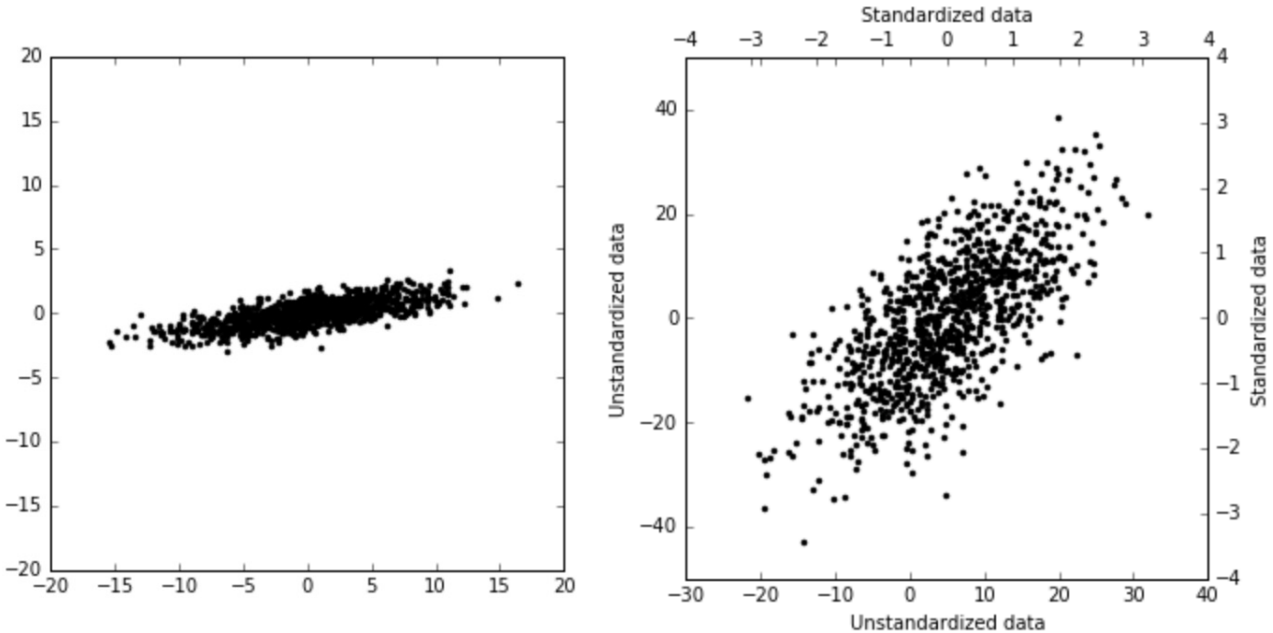
Biasing the "Proportion of Variance Explained" *Direction*

The eigenvalues point towards the most correlated axes, e.g.

$$\begin{bmatrix} 1 & 0.9 & 0.1 & 0.1 \\ 0.9 & 1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1 & 0.9 \\ 0.1 & 0.1 & 0.9 & 1 \end{bmatrix} \quad v_1 = 1/\sqrt{\frac{1}{4}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad v_2 = 1/\sqrt{\frac{1}{4}} \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & .6 & .5 \\ .6 & 1 & .4 \\ .5 & .4 & 1 \end{bmatrix} \quad v_1 = \begin{bmatrix} 0.61 \\ 0.58 \\ 0.54 \end{bmatrix}$$

The eigenvalues further point towards high variance axes, e.g.



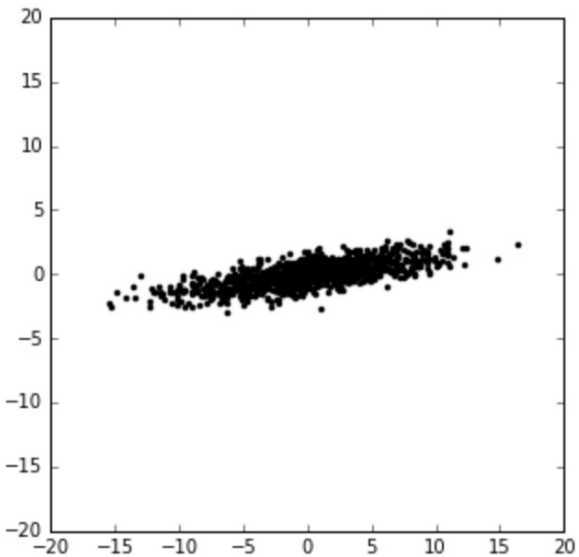
Biasing the "Proportion of Variance Explained" *Direction*

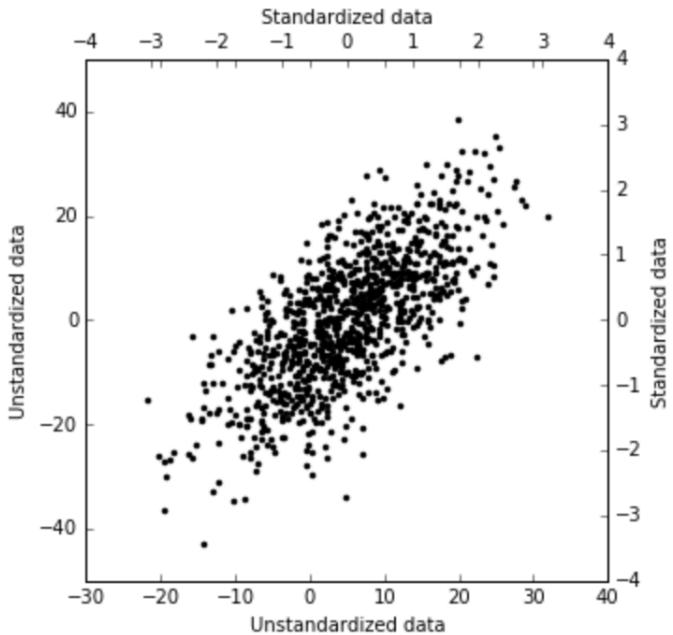
The eigenvalues point towards the most correlated axes, e.g.

$$\begin{bmatrix} 1 & 0.9 & 0.1 & 0.1 \\ 0.9 & 1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1 & 0.9 \\ 0.1 & 0.1 & 0.9 & 1 \end{bmatrix} \quad v_1 = 1/\sqrt{\frac{1}{4}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad v_2 = 1/\sqrt{\frac{1}{4}} \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & .6 & .5 \\ .6 & 1 & .4 \\ .5 & .4 & 1 \end{bmatrix} \quad v_1 = \begin{bmatrix} 0.61 \\ 0.58 \\ 0.54 \end{bmatrix}$$

The eigenvalues further point towards high variance axes, e.g.





When is $\sum_{j=1}^{p} \lambda_j = p$?

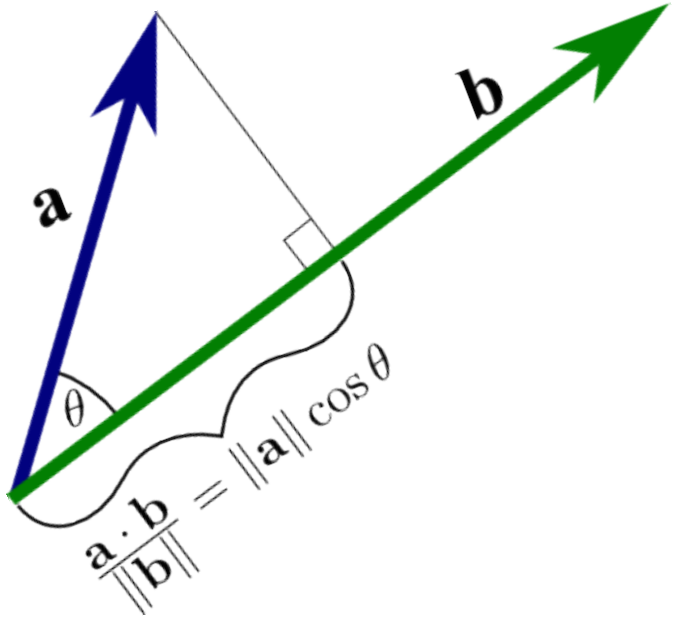
Mapping onto the directions of *"Greatest Variation"*

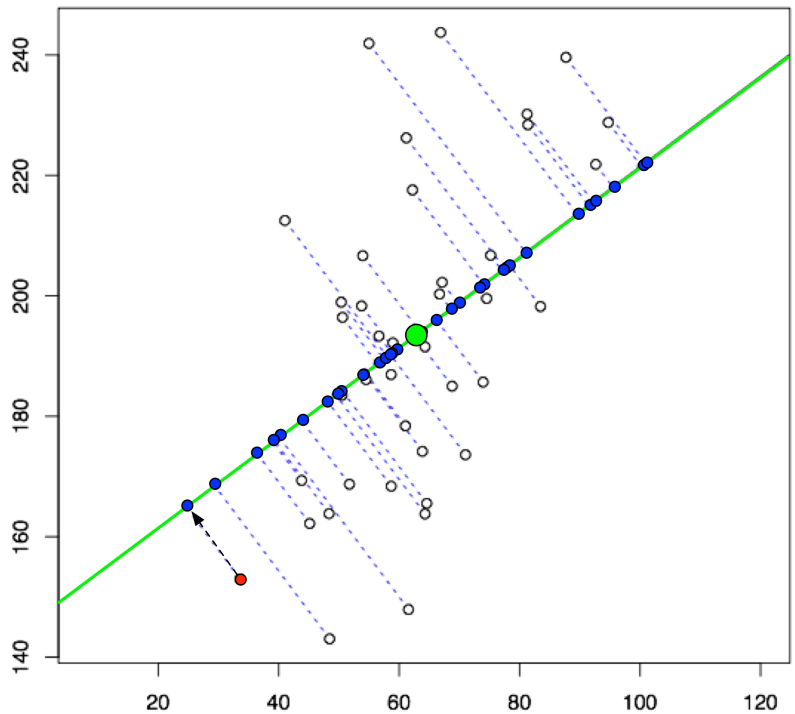
The orthogonally projection of X_i onto v_j is

$$z_{ij} = \frac{v_j}{||v_j||} \cdot X_i$$

For unit *eigenvectors* v_j, this is simply the dot product

$$z_{ij} = v_j \cdot X_i$$





Principal Components Analysis (PCA)

- PCA projects $\boldsymbol{X}$ onto the $\boldsymbol{X}$'s covariance matrix eigenvectors
- An eigenvector projection is called a principal component (PC)
- The 1^{st} *PC* is the projection onto the eigenvector with the *largest eigenvalue*, the 2^{nd} *PC* is the projection onto the eigenvector with the *second largest eigenvalue*, etc.

Principal Components Analysis (PCA)

- PCA projects $\boldsymbol{X}$ onto the $\boldsymbol{X}$'s covariance matrix eigenvectors
- An eigenvector projection is called a principal component (PC)
- The 1^{st} *PC* is the projection onto the eigenvector with the *largest eigenvalue*, the 2^{nd} *PC* is the projection onto the eigenvector with the *second largest eigenvalue*, etc.
- The *Principal Components* $\boldsymbol{z}$ are an *orthogonal rotation* of $\boldsymbol{X}$

Feature	X_1	X_2	$\cdots$	X_p
n=1				
2				
$\vdots$				
n				

$\xrightarrow{\boldsymbol{\Sigma}}$

PC	z_1	z_2	$\cdots$	z_p
n=1				
2				
$\vdots$				
n				

Principal Components Analysis (PCA)

- PCA projects $\boldsymbol{X}$ onto the $\boldsymbol{X}$'s covariance matrix eigenvectors
- An eigenvector projection is called a principal component (PC)
- The 1^{st} *PC* is the projection onto the eigenvector with the *largest eigenvalue*, the 2^{nd} *PC* is the projection onto the eigenvector with the *second largest eigenvalue*, etc.
- The *Principal Components* $\boldsymbol{z}$ are an *orthogonal rotation* of $\boldsymbol{X}$

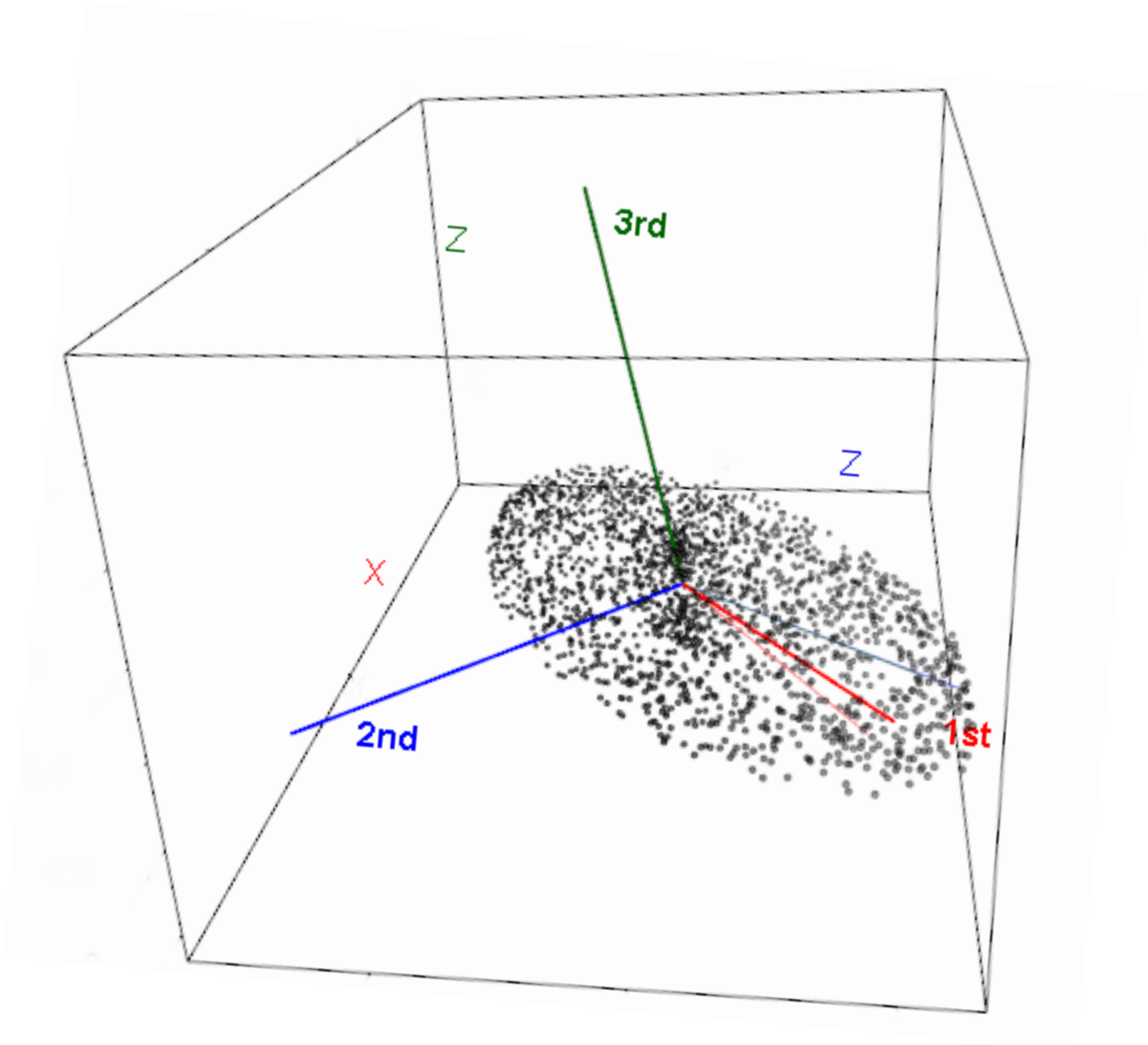
Feature	X_1	X_2	$\cdots$	X_p
n=1				
2				
$\vdots$				
n				

$\xrightarrow{\boldsymbol{\Sigma}}$

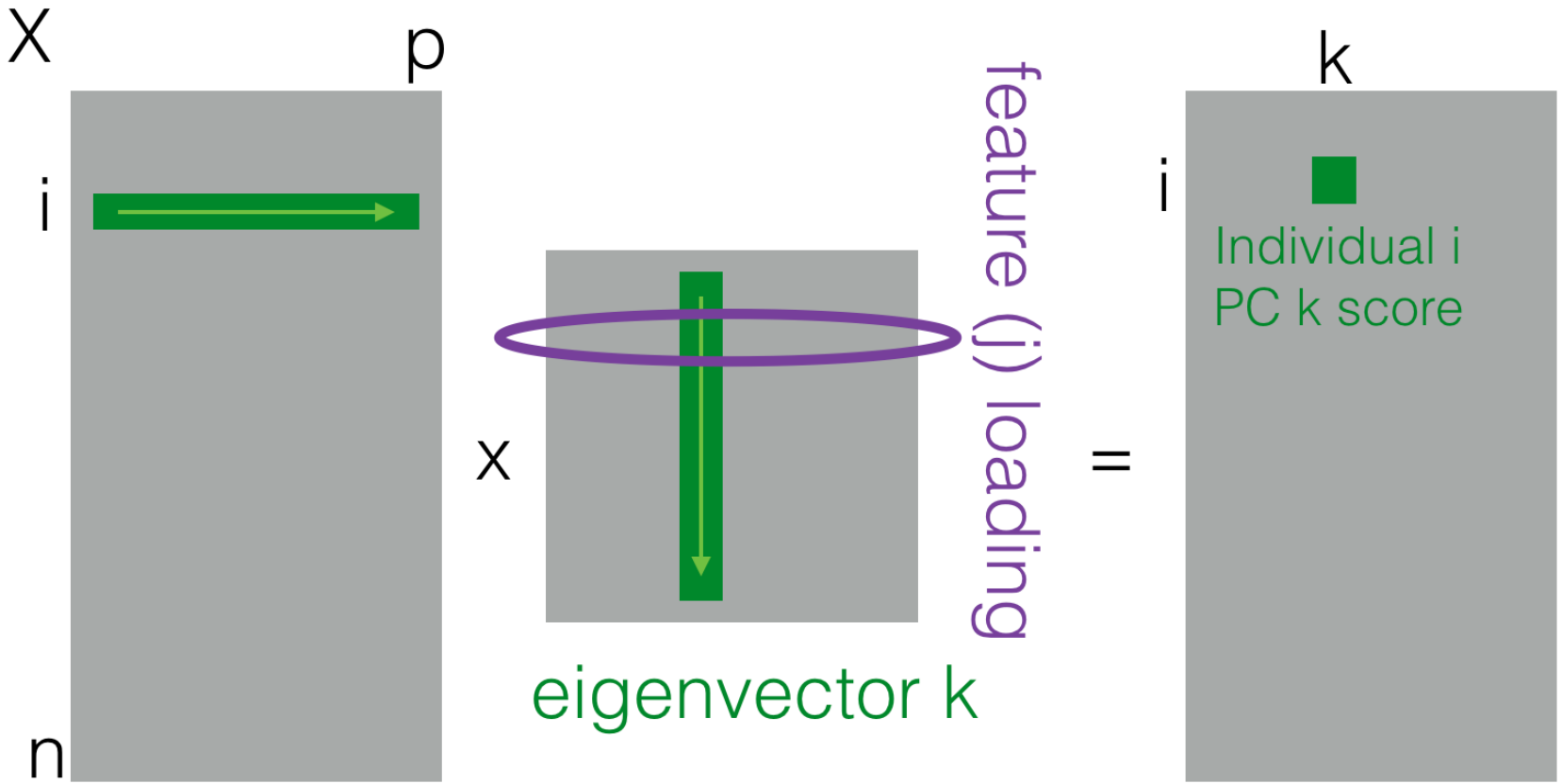
PC	z_1	z_2	$\cdots$	z_p
n=1				
2				
$\vdots$				
n				

- PCA routinely centers and scales $X_{\cdot j}$ in order to capture correlation between features rather than feature scales

PCA: **Transforms Features To Uncorrelated Features**



Mechanics and Terms



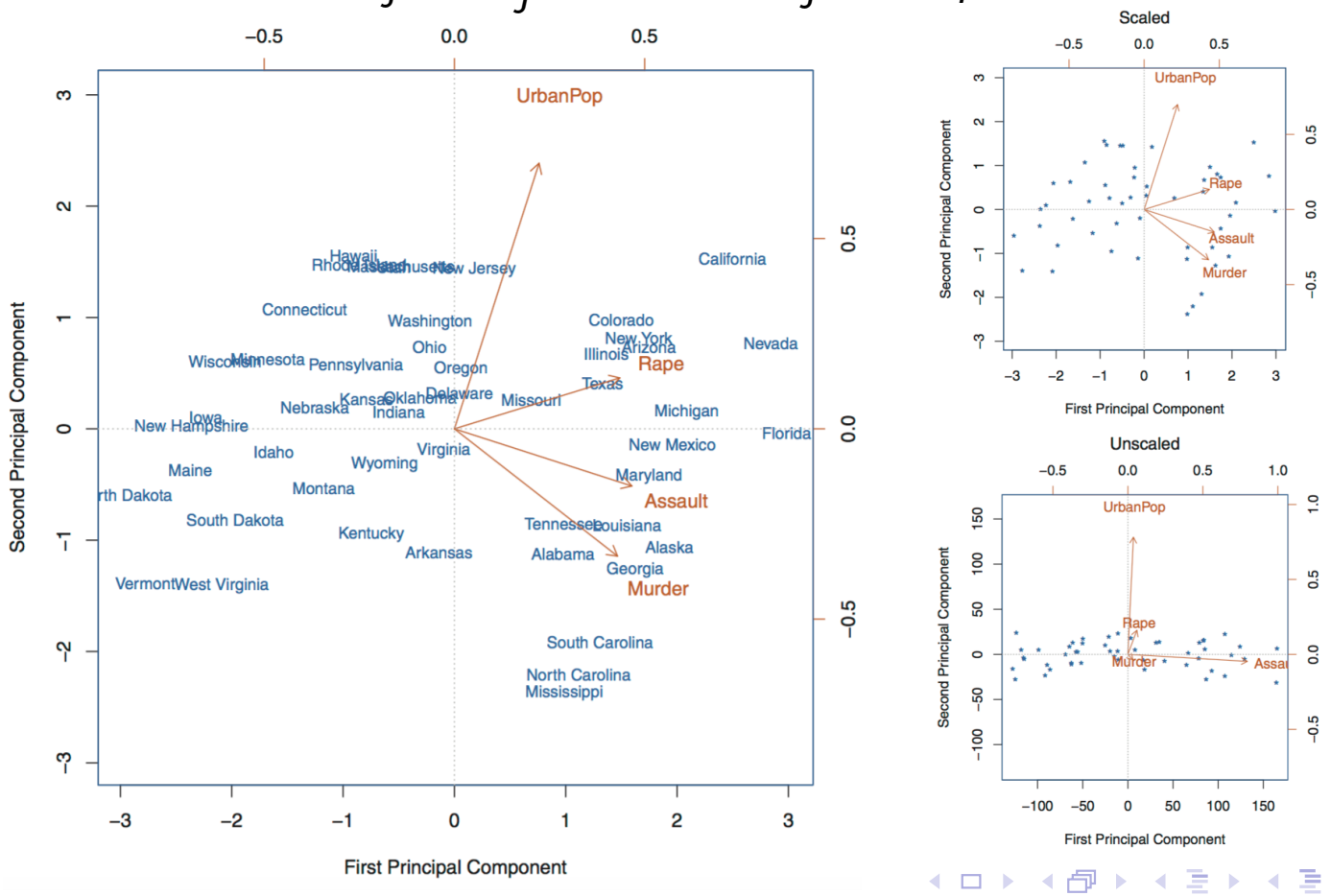
Biplots: *look at the axes*

- Vector: k^{th} versus k'^{th} eigenvector's j^{th} feature loading:

$$v_{kj} \text{ versus } v_{k'j}$$

- Point: observation i's j^{th} versus j'^{th} PC scores:

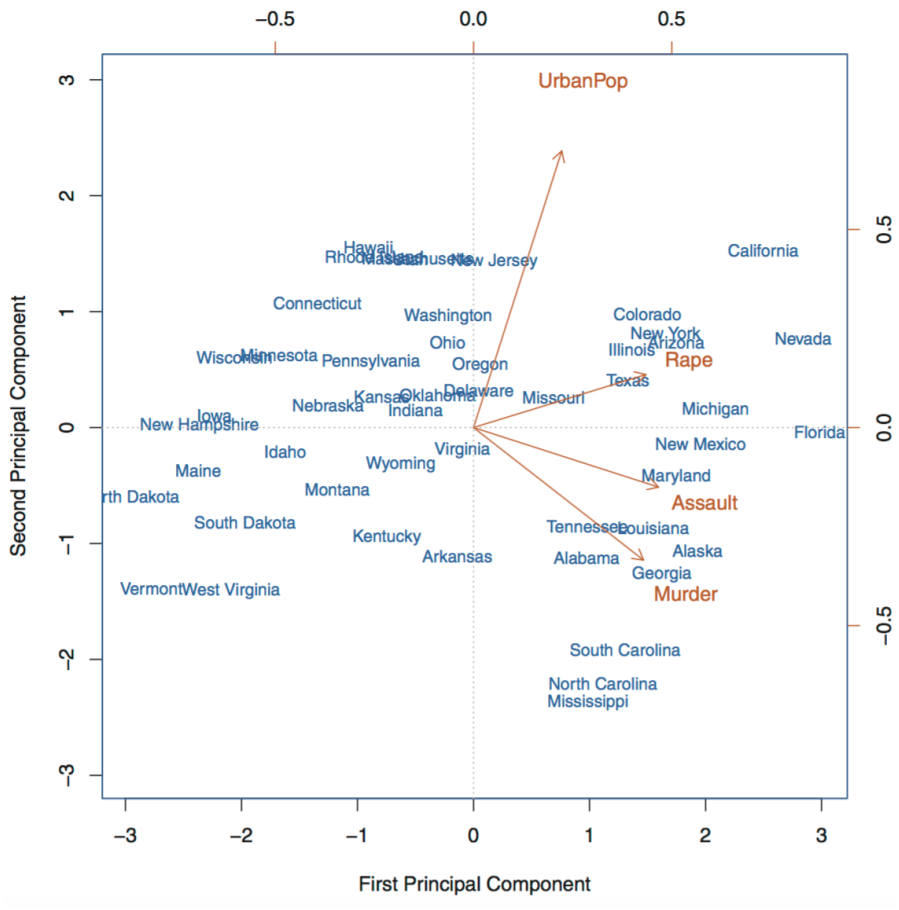
$$z_{ji} = v'_j X_i \text{ versus } z_{j'i} = v'_{i'} X_i$$

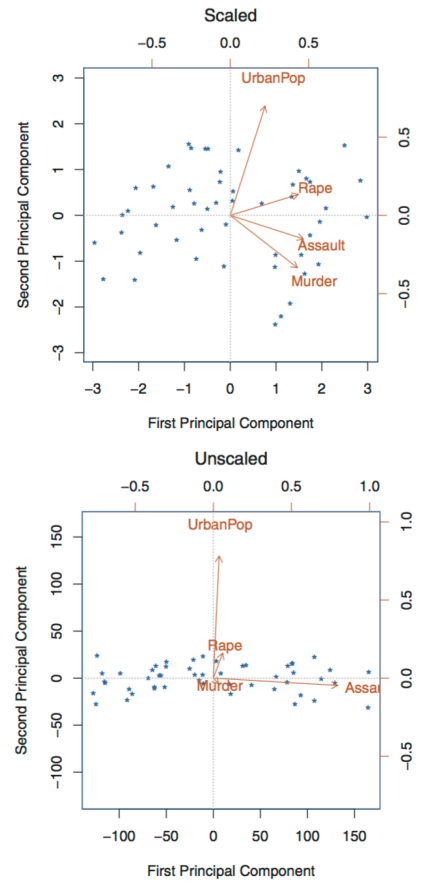


Biplots: *look at the axes*

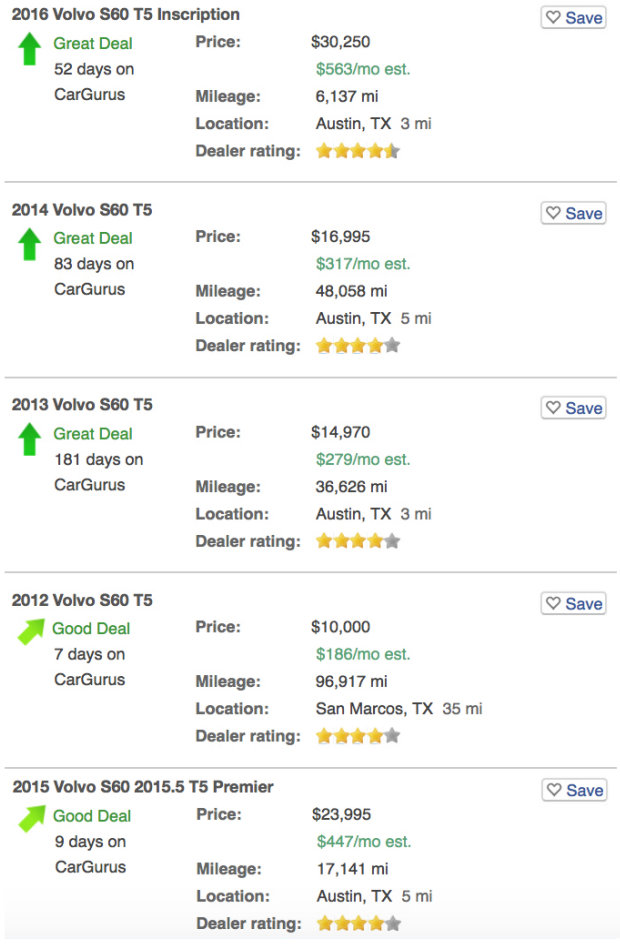
- Vector: which features the PCs point towards

- Point: which data points are driven by which PCs





Dimensionality Reduction?



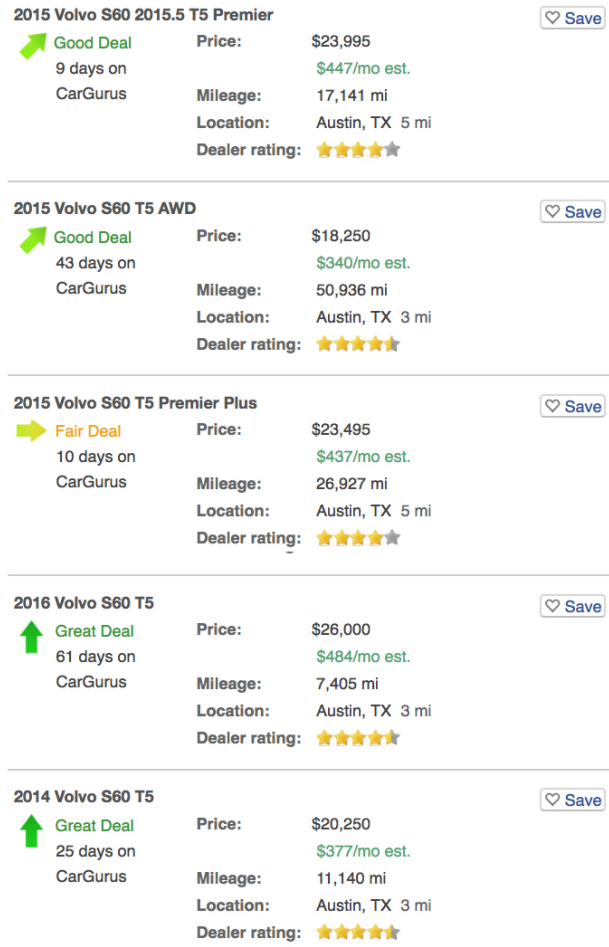
2016 Volvo S60 T5 Inscription — ♡ Save
Great Deal
52 days on CarGurus
Price: $30,250
$563/mo est.
Mileage: 6,137 mi
Location: Austin, TX 3 mi
Dealer rating:

2014 Volvo S60 T5 — ♡ Save
Great Deal
83 days on CarGurus
Price: $16,995
$317/mo est.
Mileage: 48,058 mi
Location: Austin, TX 5 mi
Dealer rating:

2013 Volvo S60 T5 — ♡ Save
Great Deal
181 days on CarGurus
Price: $14,970
$279/mo est.
Mileage: 36,626 mi
Location: Austin, TX 3 mi
Dealer rating:

2012 Volvo S60 T5 — ♡ Save
Good Deal
7 days on CarGurus
Price: $10,000
$186/mo est.
Mileage: 96,917 mi
Location: San Marcos, TX 35 mi
Dealer rating:

2015 Volvo S60 2015.5 T5 Premier — ♡ Save
Good Deal
9 days on CarGurus
Price: $23,995
$447/mo est.
Mileage: 17,141 mi
Location: Austin, TX 5 mi
Dealer rating:



2015 Volvo S60 2015.5 T5 Premier — ♡ Save
Good Deal
9 days on CarGurus
Price: $23,995
$447/mo est.
Mileage: 17,141 mi
Location: Austin, TX 5 mi
Dealer rating:

2015 Volvo S60 T5 AWD — ♡ Save
Good Deal
43 days on CarGurus
Price: $18,250
$340/mo est.
Mileage: 50,936 mi
Location: Austin, TX 3 mi
Dealer rating:

2015 Volvo S60 T5 Premier Plus — ♡ Save
Fair Deal
10 days on CarGurus
Price: $23,495
$437/mo est.
Mileage: 26,927 mi
Location: Austin, TX 5 mi
Dealer rating:

2016 Volvo S60 T5 — ♡ Save
Great Deal
61 days on CarGurus
Price: $26,000
$484/mo est.
Mileage: 7,405 mi
Location: Austin, TX 3 mi
Dealer rating:

2014 Volvo S60 T5 — ♡ Save
Great Deal
25 days on CarGurus
Price: $20,250
$377/mo est.
Mileage: 11,140 mi
Location: Austin, TX 3 mi
Dealer rating:

Dimensionality Reduction?

2016 Volvo S60 T5 Inscription ♡ Save

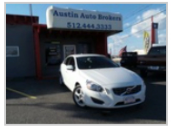
Great Deal
52 days on CarGurus

Price: $30,250
$563/mo est.
Mileage: 6,137 mi
Location: Austin, TX 3 mi
Dealer rating:

2014 Volvo S60 T5 ♡ Save

Great Deal
83 days on CarGurus

Price: $16,995
$317/mo est.
Mileage: 48,058 mi
Location: Austin, TX 5 mi
Dealer rating:



2013 Volvo S60 T5 ♡ Save

Great Deal
181 days on CarGurus

Price: $14,970
$279/mo est.
Mileage: 36,626 mi
Location: Austin, TX 3 mi
Dealer rating:

2012 Volvo S60 T5 ♡ Save

Good Deal
7 days on CarGurus

Price: $10,000
$186/mo est.
Mileage: 96,917 mi
Location: San Marcos, TX 35 mi
Dealer rating:

2015 Volvo S60 2015.5 T5 Premier ♡ Save

Good Deal
9 days on CarGurus

Price: $23,995
$447/mo est.
Mileage: 17,141 mi
Location: Austin, TX 5 mi
Dealer rating:

2015 Volvo S60 2015.5 T5 Premier ♡ Save

Good Deal
9 days on CarGurus

Price: $23,995
$447/mo est.
Mileage: 17,141 mi
Location: Austin, TX 5 mi
Dealer rating:

2015 Volvo S60 T5 AWD ♡ Save

Good Deal
43 days on CarGurus

Price: $18,250
$340/mo est.
Mileage: 50,936 mi
Location: Austin, TX 3 mi
Dealer rating:

2015 Volvo S60 T5 Premier Plus ♡ Save

Fair Deal
10 days on CarGurus

Price: $23,495
$437/mo est.
Mileage: 26,927 mi
Location: Austin, TX 5 mi
Dealer rating:

2016 Volvo S60 T5 ♡ Save

Great Deal
61 days on CarGurus

Price: $26,000
$484/mo est.
Mileage: 7,405 mi
Location: Austin, TX 3 mi
Dealer rating:

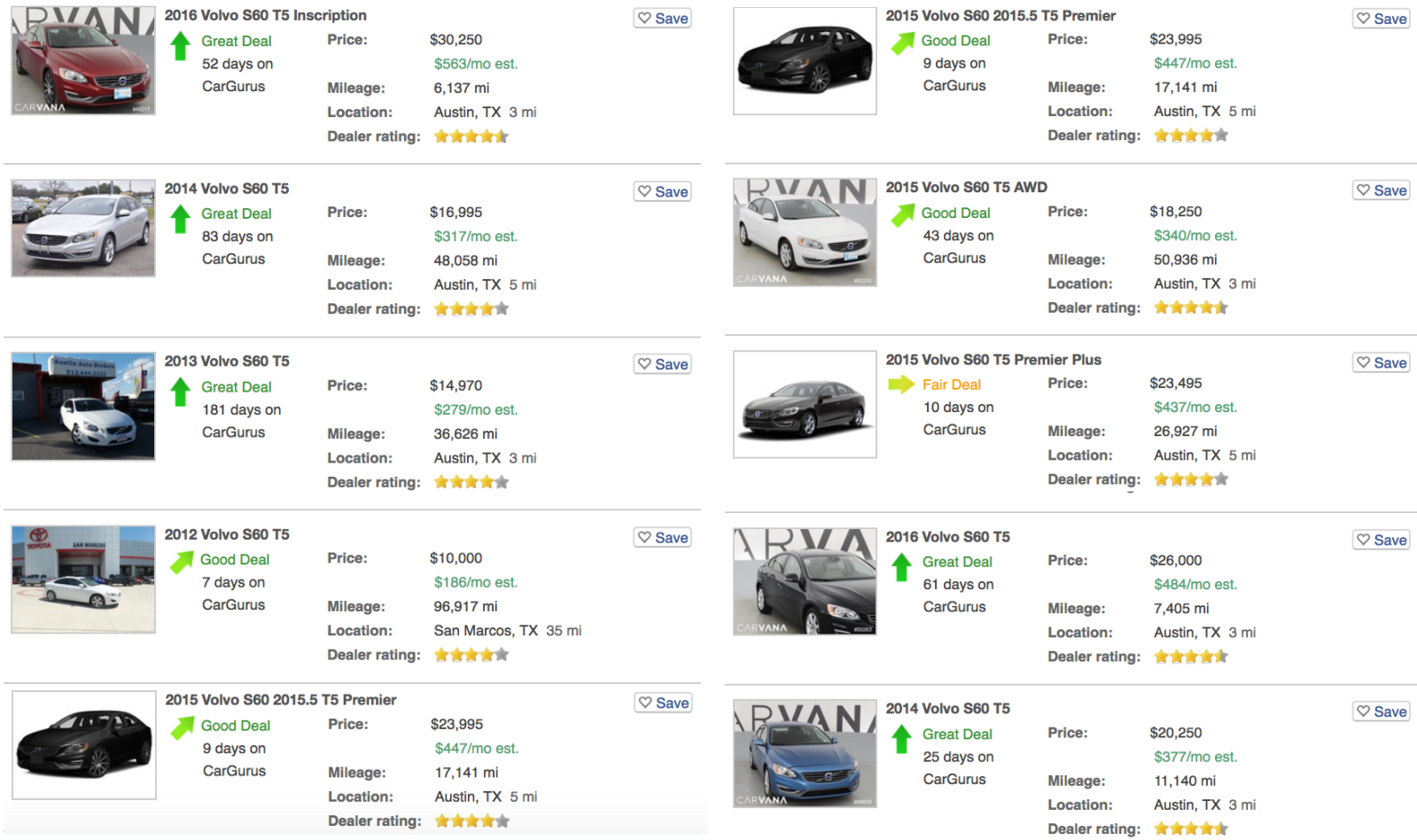
2014 Volvo S60 T5 ♡ Save

Great Deal
25 days on CarGurus

Price: $20,250
$377/mo est.
Mileage: 11,140 mi
Location: Austin, TX 3 mi
Dealer rating:

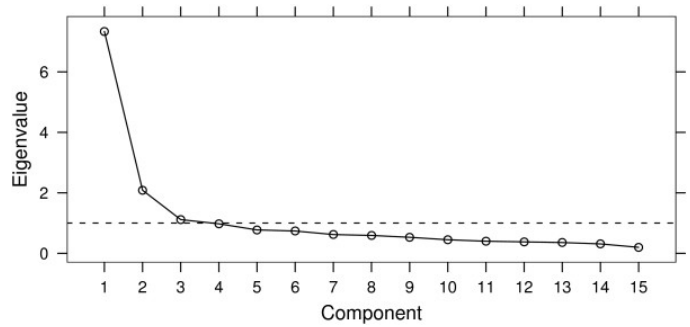
Dimensionality Reduction?

How does e.g. KNN feel about this many dimensions? Curse of...



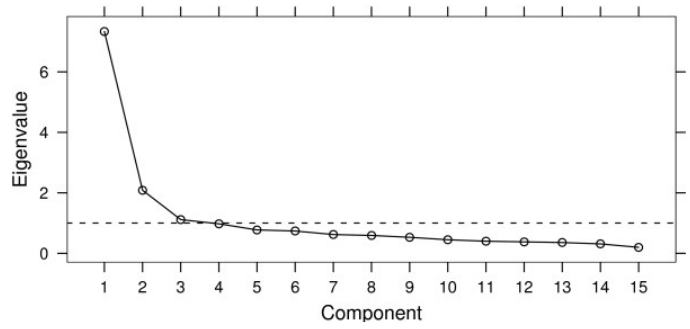
Dimensionality Reduction

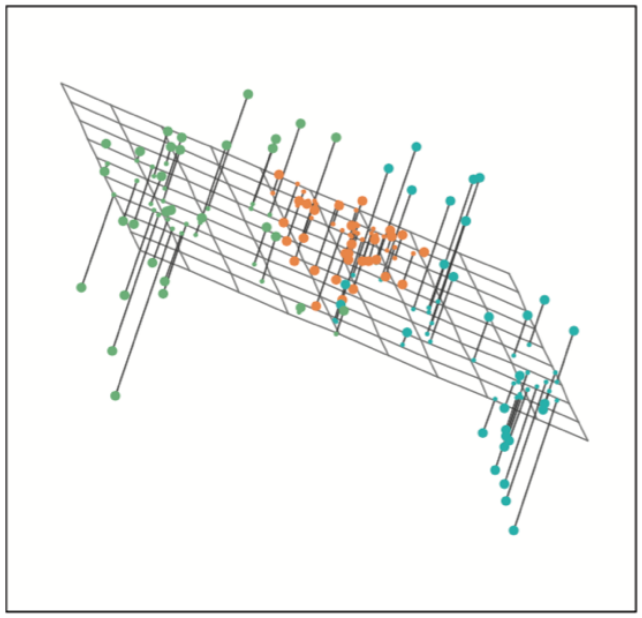
$$\underset{p\times 1}{X_i} \approx \sum_{j=1}^{q<p} \underset{p\times 1}{v_j}\, z_{ij}$$

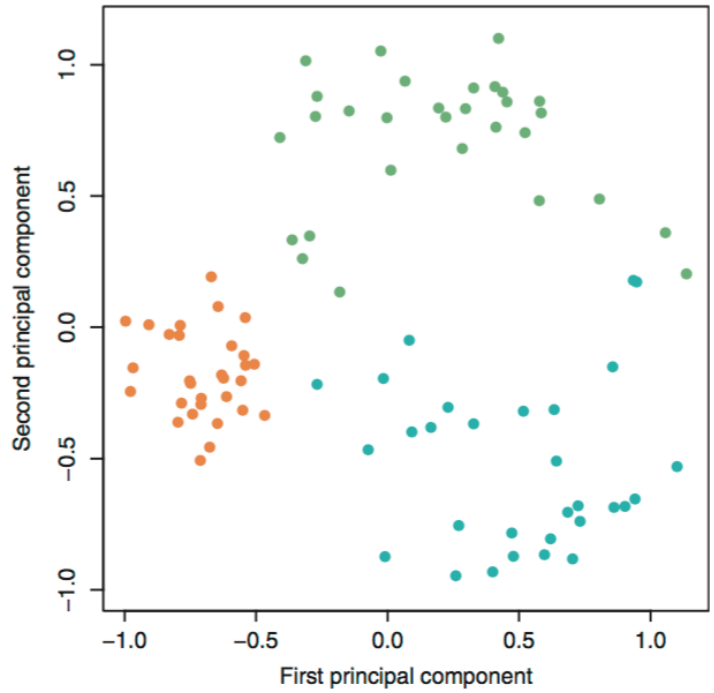


Dimensionality Reduction

$$\underset{p\times 1}{X_i} \approx \sum_{j=1}^{q<p} \underset{p\times 1}{v_j}\, z_{ij}$$

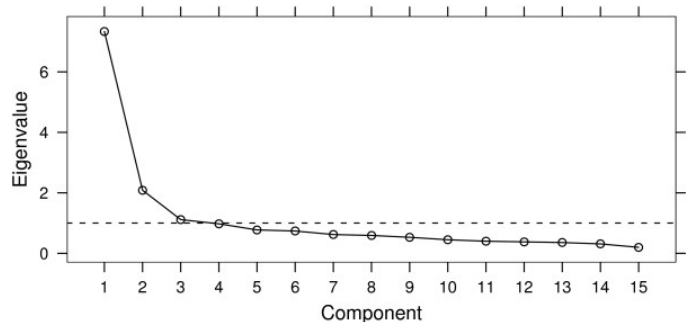


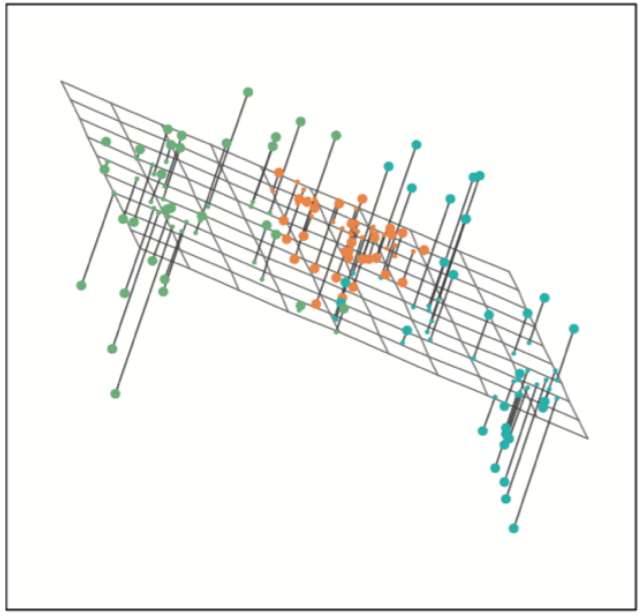


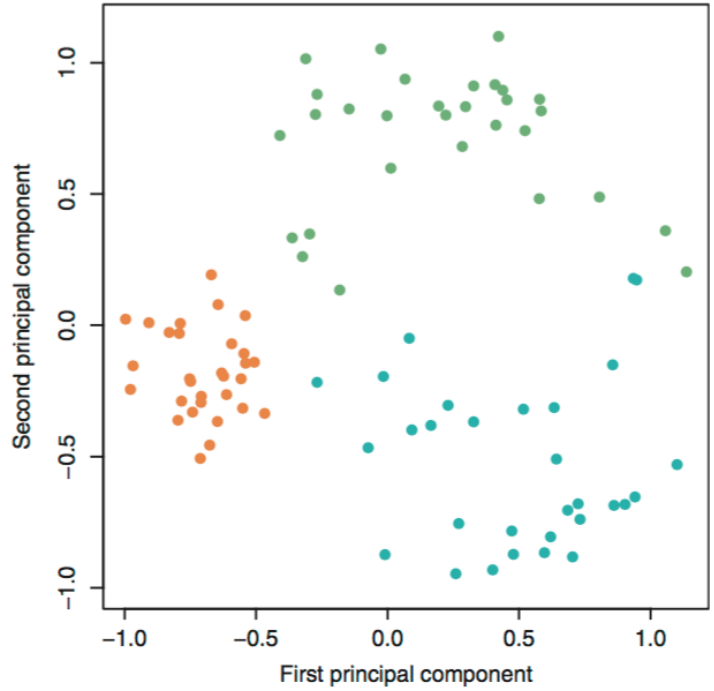


Dimensionality Reduction

$$\underset{p\times 1}{X_i} \approx \sum_{j=1}^{q<p} \underset{p\times 1}{v_j}\, z_{ij}$$



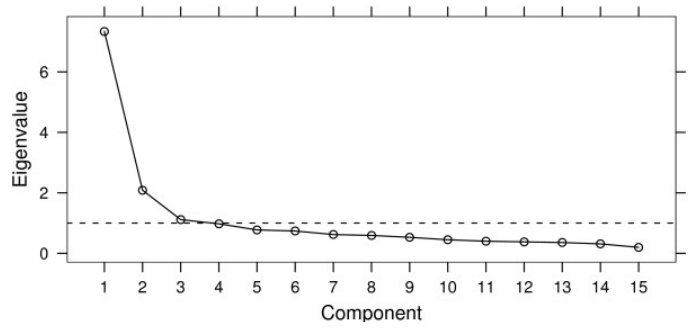


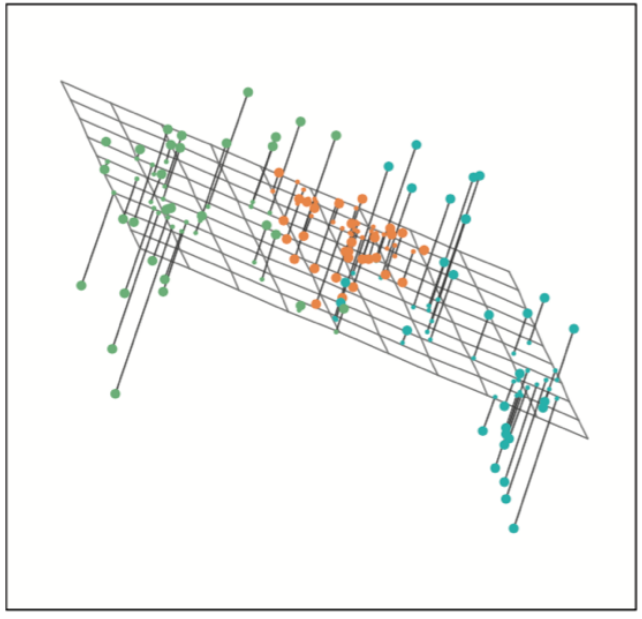


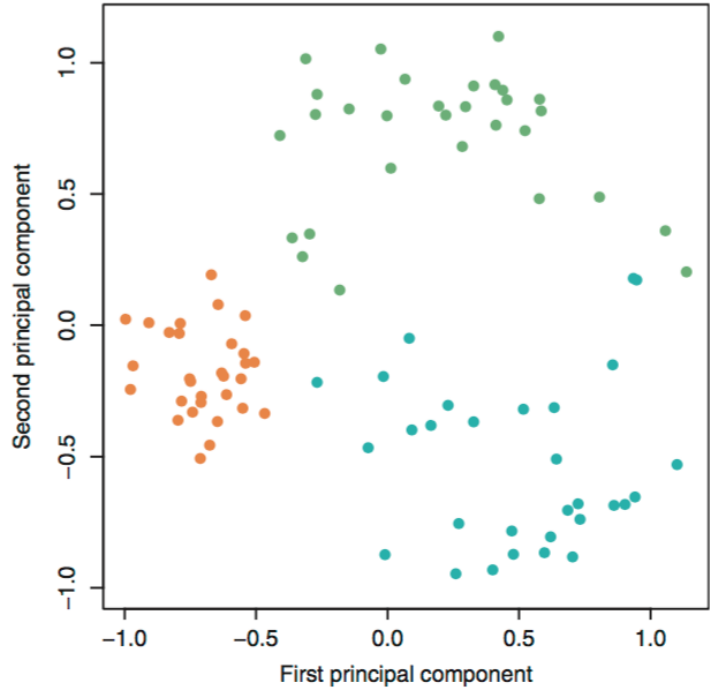
Compresses data into uncorrelated variation & maybe drops noise
(Feature Engineering?) (Curse of Dimensionality?)

Dimensionality Reduction

$$\underset{p\times 1}{X_i} \approx \sum_{j=1}^{q<p} \underset{p\times 1}{v_j}\, z_{ij}$$







Compresses data into uncorrelated variation & maybe drops noise

(Feature Engineering?) (Curse of Dimensionality?)

The number of *nonzero* λ_j's equals the rank of the covariance matrix $\boldsymbol{X}^T\boldsymbol{X}$ which itself equals the rank of the data matrix $\boldsymbol{X}$

Dimensionality Reduction

Feature	X_1	X_2	$\cdots$	X_p
n=1				
2				
$\vdots$				
n				

$\xrightarrow{\boldsymbol{\Sigma}}$

PC	z_1	z_2	$\cdots$	z_p
n=1				
2				
$\vdots$				
n				

Dimensionality Reduction

Feature	X_1	X_2	$\cdots$	X_p
n=1				
2				
$\vdots$				
n				

$\xrightarrow{\Sigma}$

PC	z_1	z_2	$\cdots$	z_p
n=1				
2				
$\vdots$				
n				

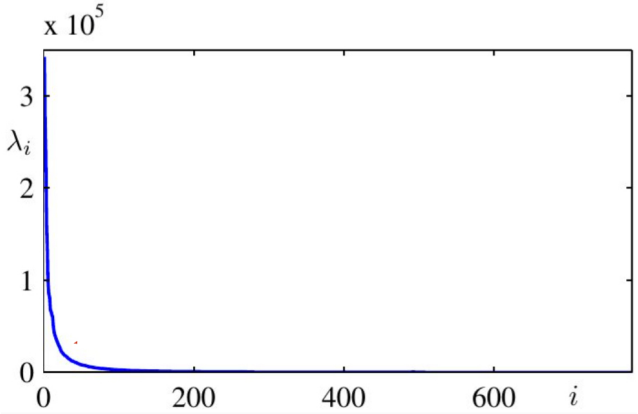
Feature	X_1	X_2	$\cdots$	X_p
n=1				
2				
$\vdots$				
n				

$\xrightarrow{\Sigma}$

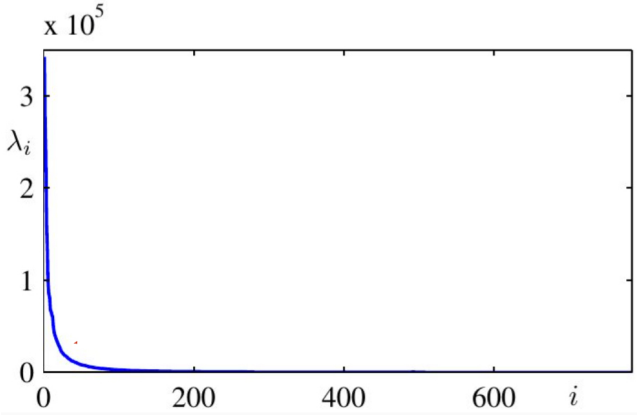
PC	z_1	$\cdots$
n=1		
2		
$\vdots$		
n		

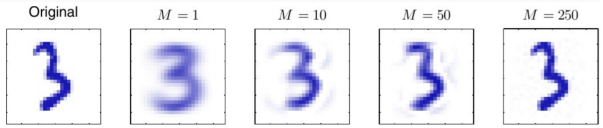
Dimensionality Reduction Examples

Dimensionality Reduction Examples

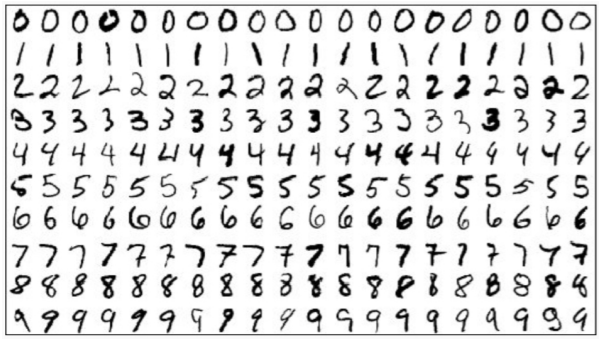


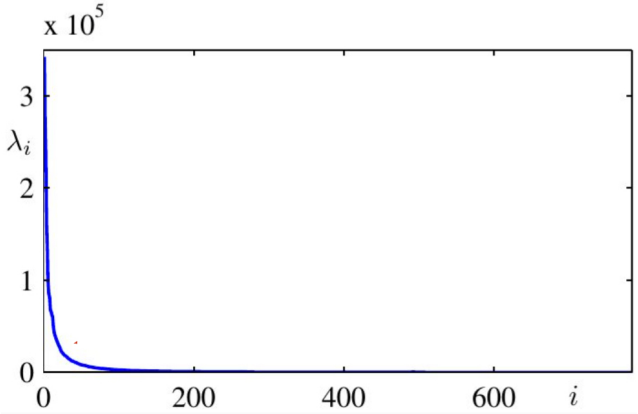
Dimensionality Reduction Examples

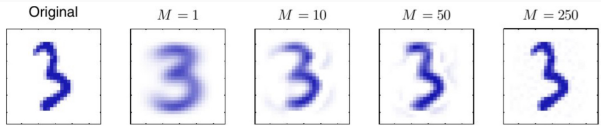


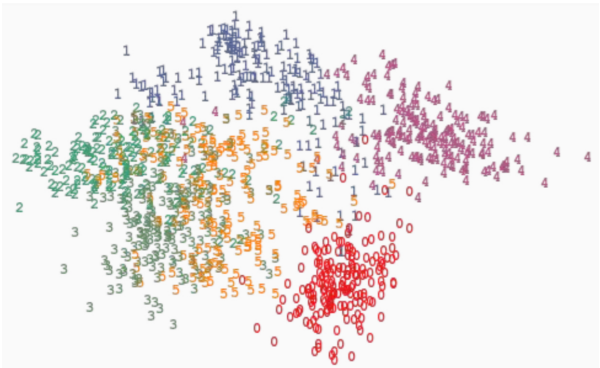


Dimensionality Reduction Examples

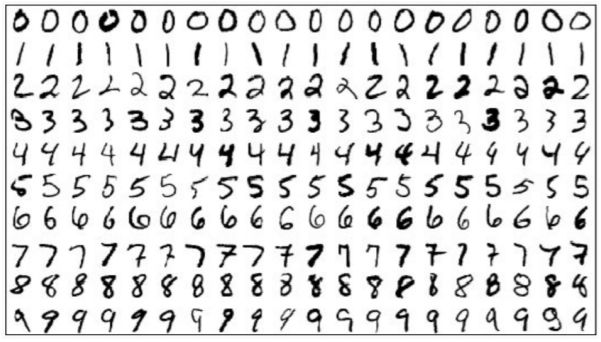


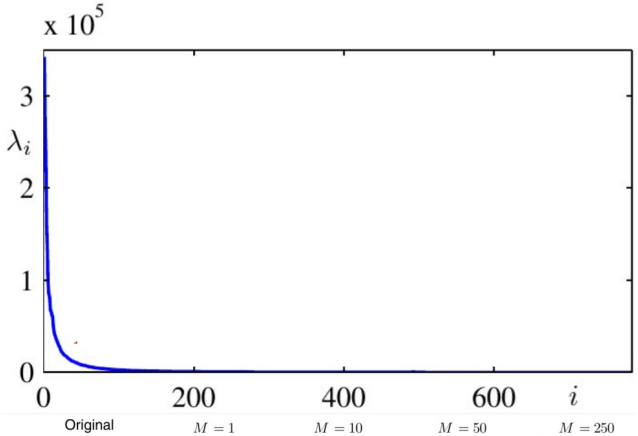


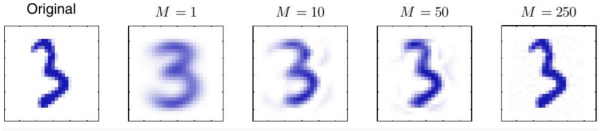




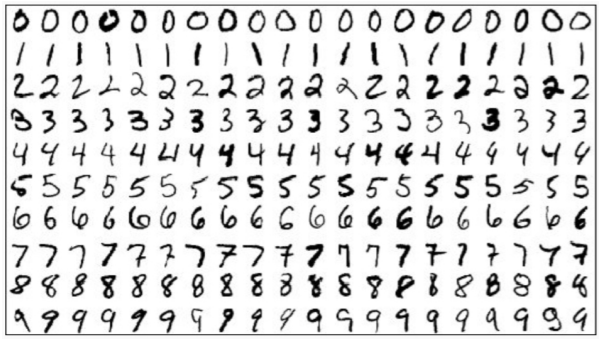
Dimensionality Reduction Examples

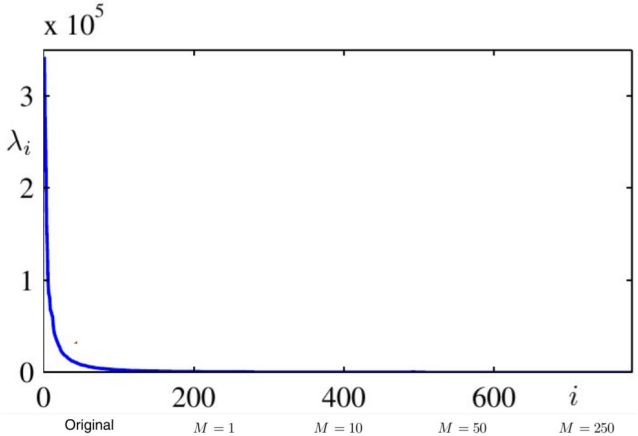


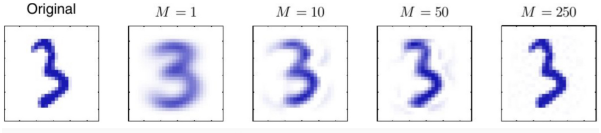


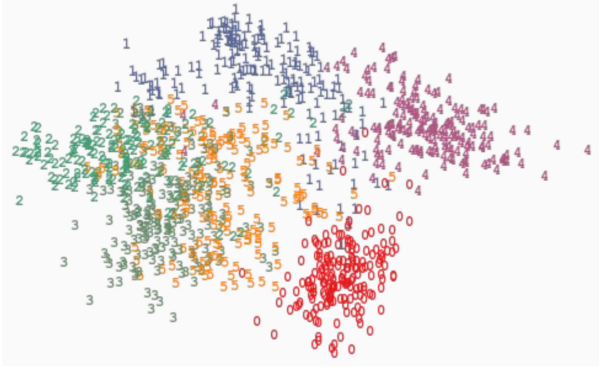


Dimensionality Reduction Examples

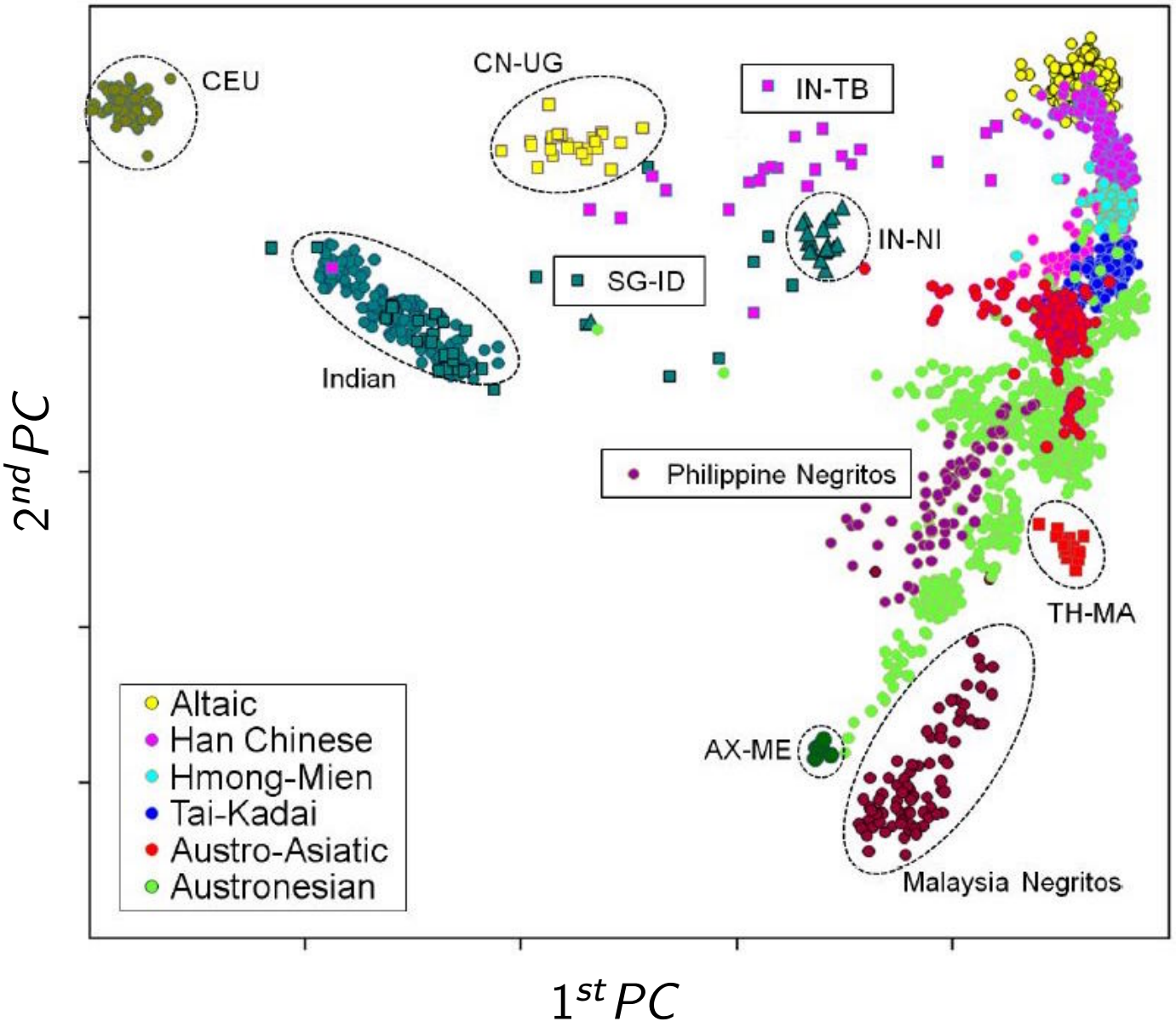








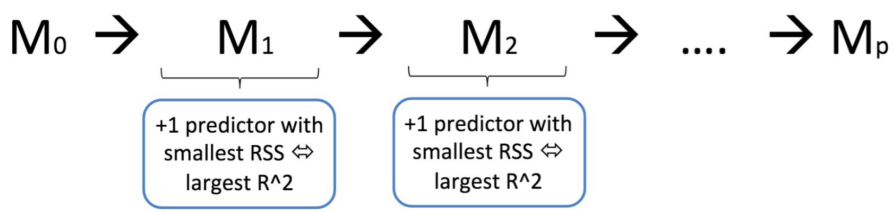
Visualization (for EDA purposes)



How else could we do dimension reduction?

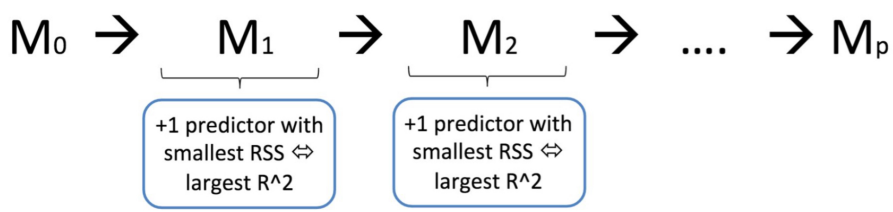
How else could we do dimension reduction?

- Stepwise Selection

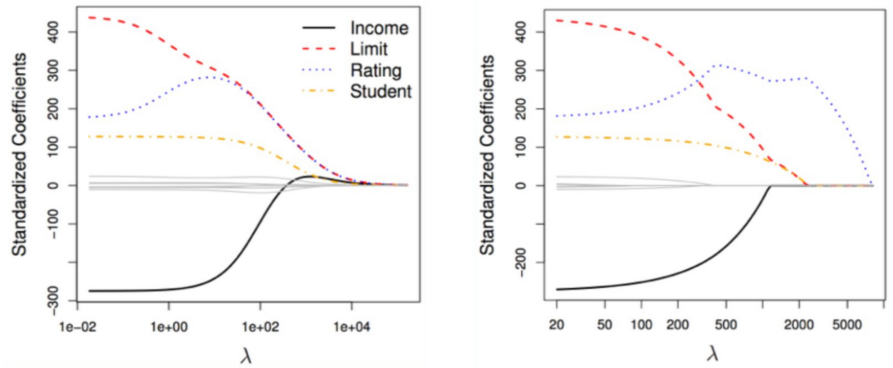


How else could we do dimension reduction?

- Stepwise Selection

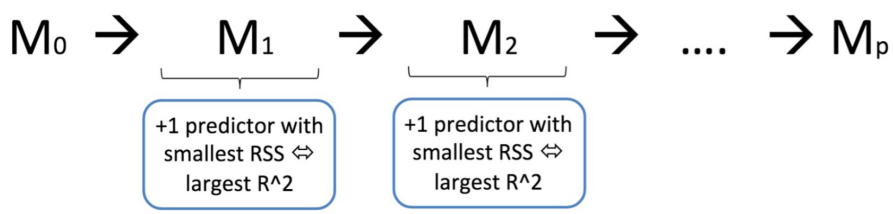


- Lasso

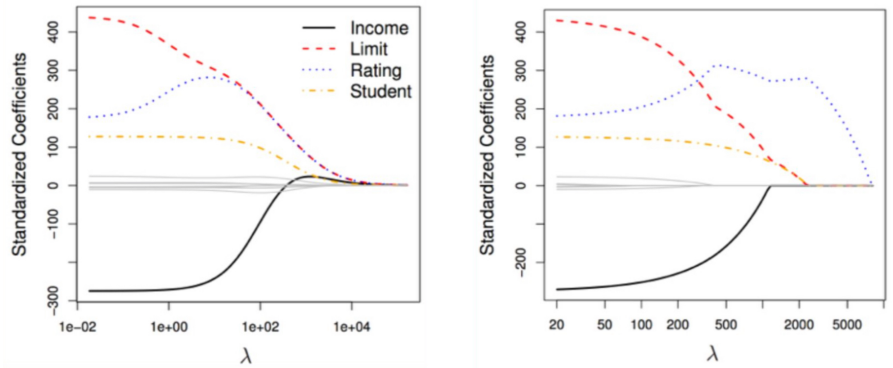


How else could we do dimension reduction?

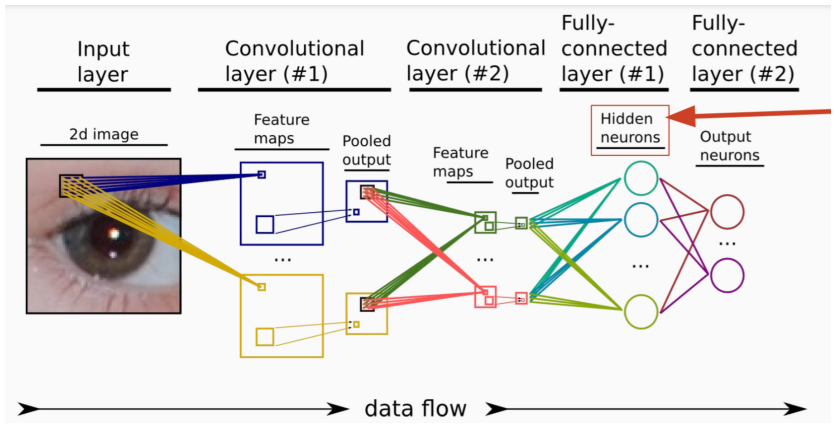
- Stepwise Selection



- Lasso

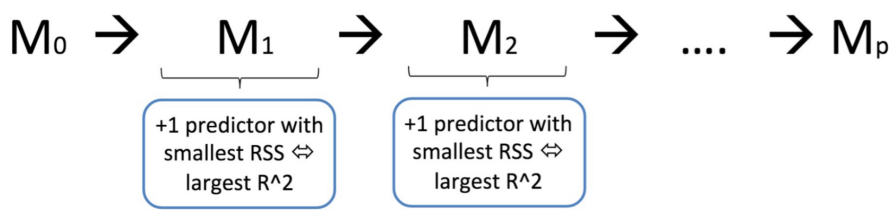


- Neural Networks

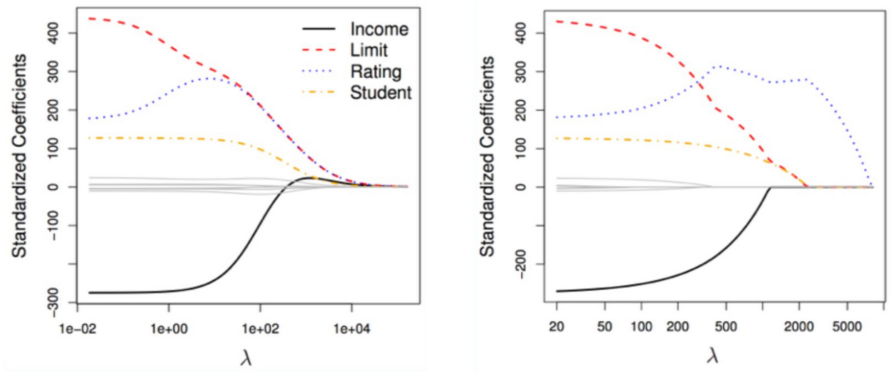


How else could we do dimension reduction?

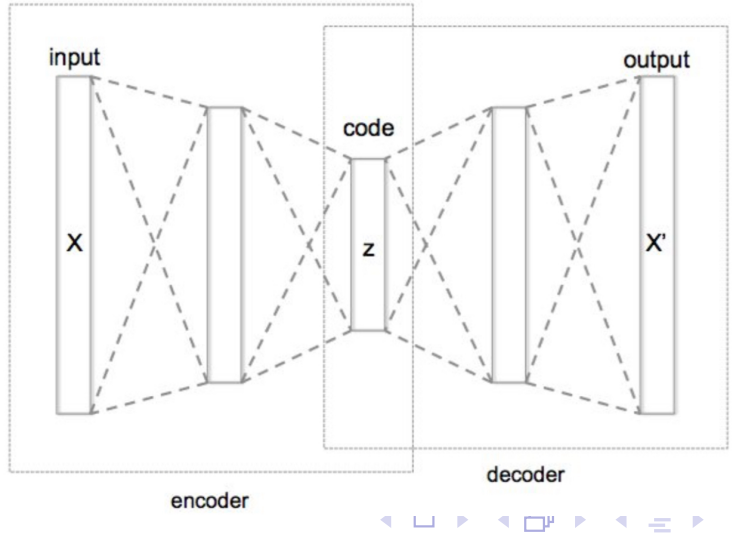
- Stepwise Selection



- Lasso

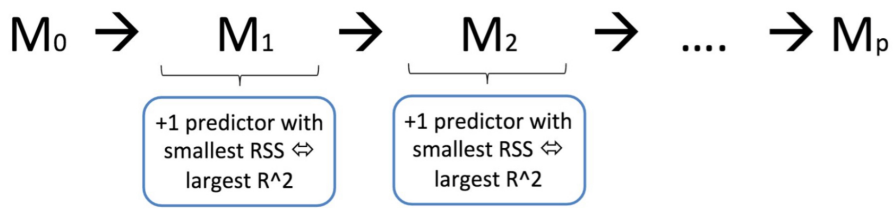


- Neural Networks

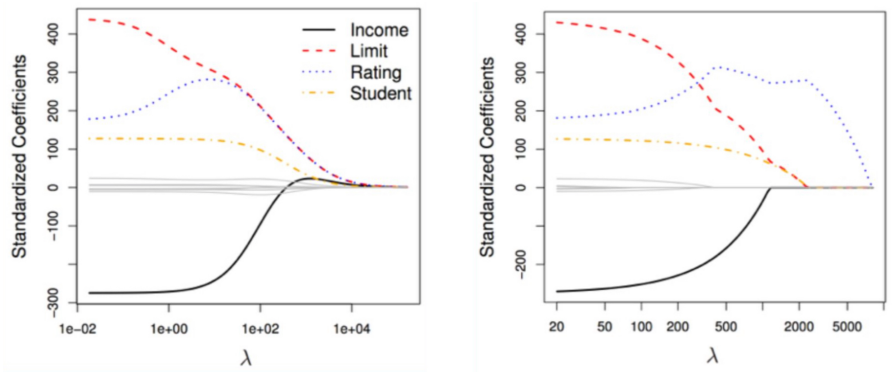


How else could we do dimension reduction?

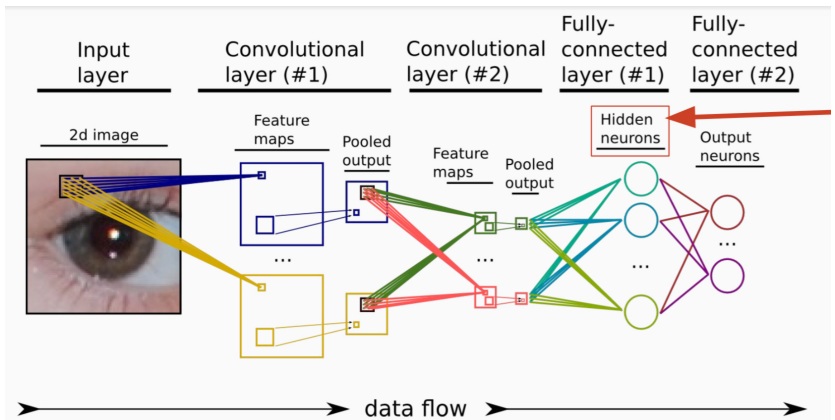
- Stepwise Selection



- Lasso

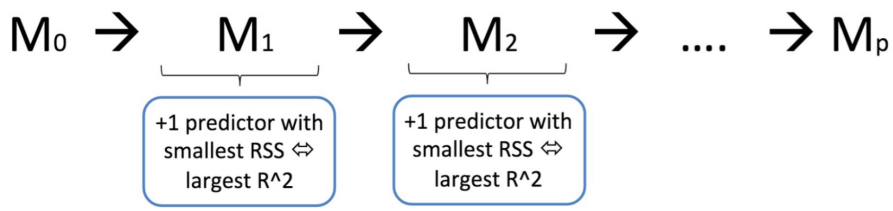


- Neural Networks

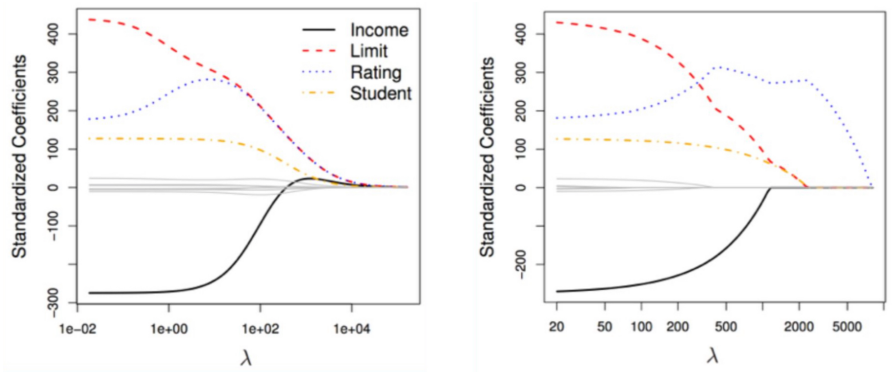


How else could we do dimension reduction?

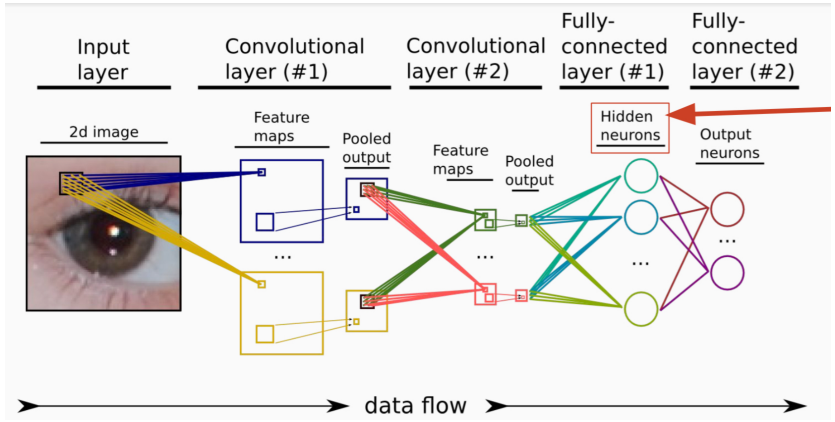
- Stepwise Selection



- Lasso



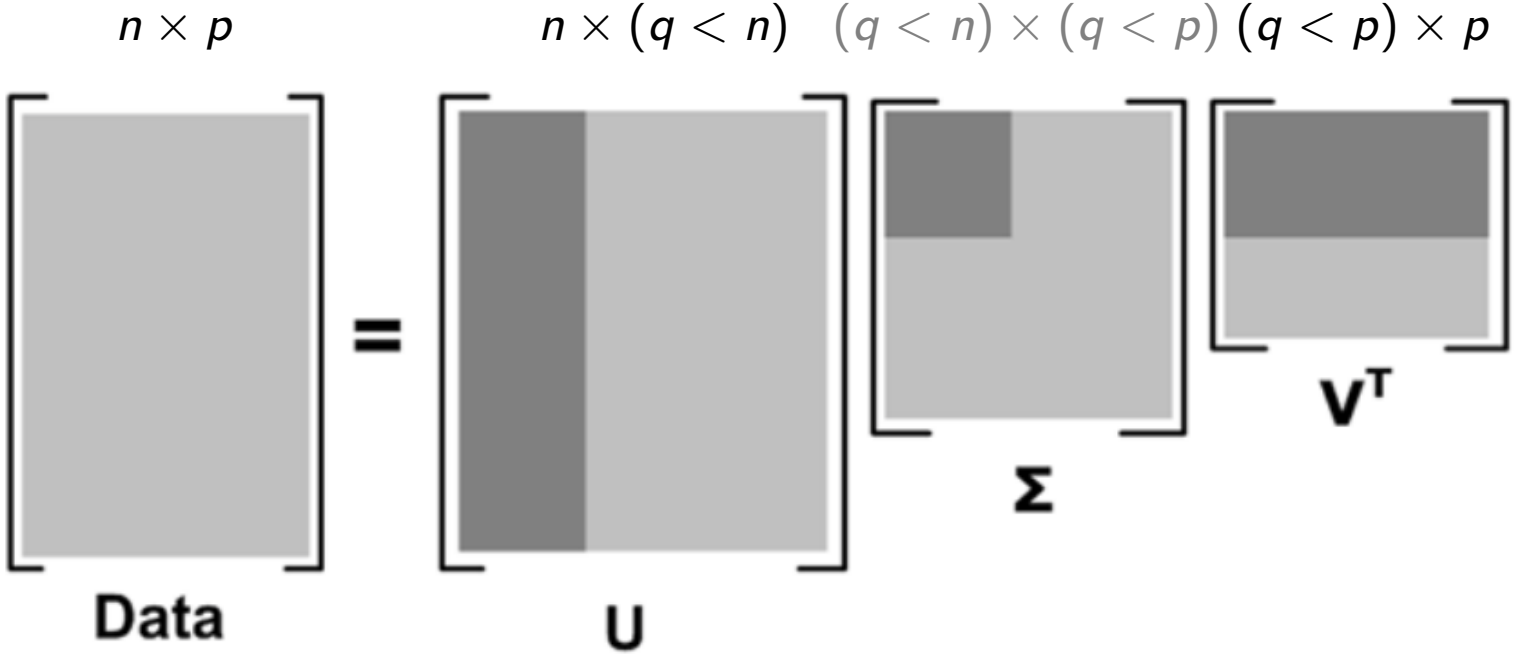
- Neural Networks



It's fun to stay at the

- PCA, eh?

Singular Value Decomposition (SVD)

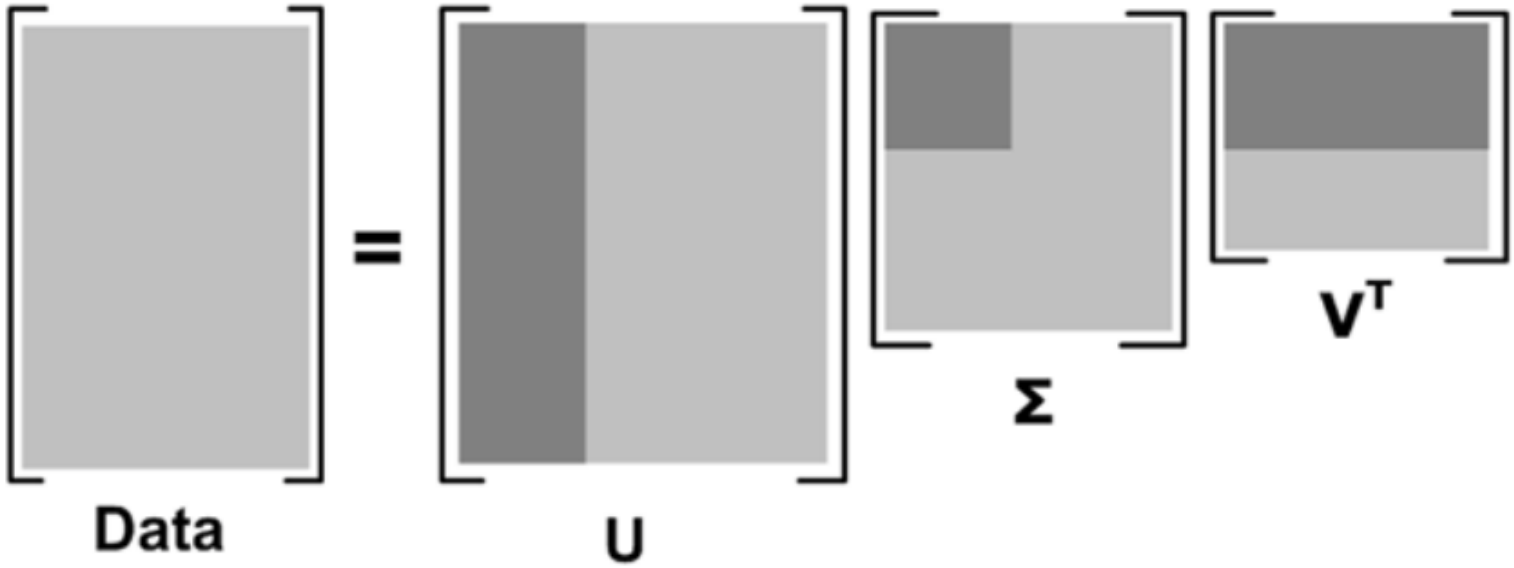


- Data: "standarized" $\boldsymbol{X}$
- U: Standardized principal component scores
- Σ: $\sqrt{eigenvalues}$ on diagonal, 0's off diagonal
- V^T: Matrix whose rows are the eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$

$$\underset{1\times 1}{X_{ij'}} \approx \sum_{j=1}^{q<p} \underset{1\times 1}{U_{ij}} \sqrt{\lambda_j} \underset{1\times 1}{V^T_{jj'}}$$

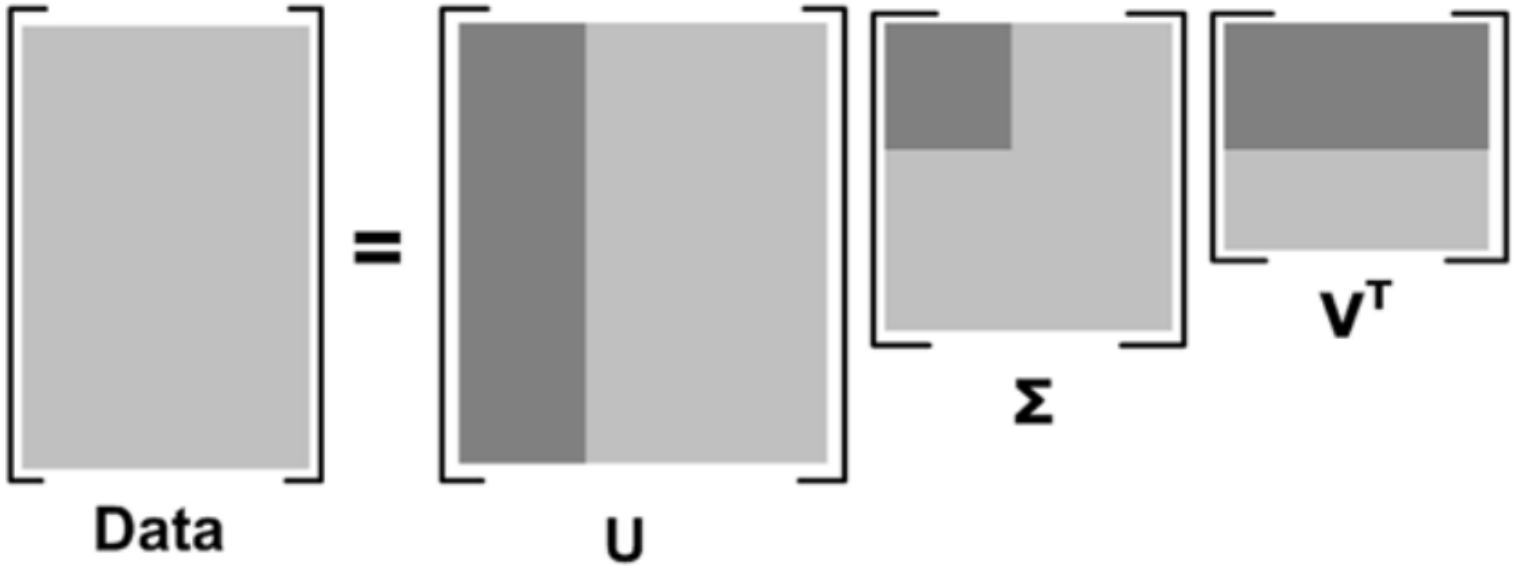
$$\underset{p\times 1}{X_i} \approx \sum_{j=1}^{q<p} \underset{p\times 1}{v_j}\, z_{ij}$$

Latent Factors



- Data is the "phoenotypes" (features)
- U is the "genes" (factors) driving "phoenotypes" (features)
- V^T is the "way" "genes" drive "phoenotypes"
- Σ controls the strength each "genes" influence

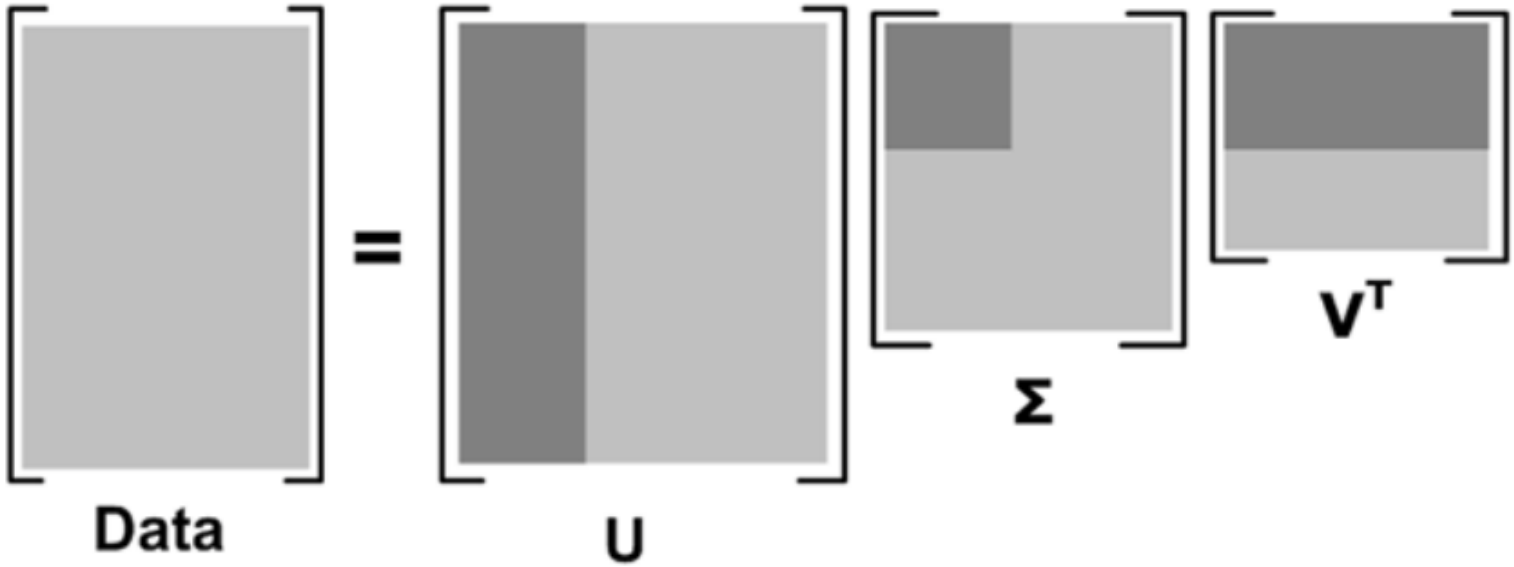
Latent Factors



- Data is the "phoenotypes" (features)
- U is the "genes" (factors) driving "phoenotypes" (features)
- V^T is the "way" "genes" drive "phoenotypes"
- Σ controls the strength each "genes" influence

- For some "high dimensional" data, e.g., audio, video, these *latent factors* are the meaningful features associating with Y's

Latent Factors



- When $p >> n$, $\boldsymbol{X}^T\boldsymbol{X}$ is huge
- SVD doesn't require computation of $\boldsymbol{X}^T\boldsymbol{X}$ (i.e., $\boldsymbol{S}$)
- SVD is a faster, more stable algorithm (than PCA)

Bonus: Factor Analysis – genes and phenotypes

$$X_{ij} - \mu_j = \sum_{k=1}^{q} \beta_{jk} F_{ik} + \epsilon_{ij}$$

	What	Index	Variable	Note
n	samples	$i = 1, \cdots, n$		
p	features/sample	$j = 1, \cdots, p$	X_{ij}	$E[X_{ij}] = \mu_j$
p	errors/sample	$j = 1, \cdots, p$	ϵ_{ij}	$\overset{iid}{\sim} N(0, \sigma_j)$
q	factors/sample	$k = 1, \cdots, q$	F_{ik}	$\overset{iid}{\sim}$ N(0,1)
$p \times q$	factor loadings		β_{jk}	Constants

So for each sample i we have p equations

$$\begin{aligned}
X_{i1} - \mu_1 &= \beta_{11} F_{i1} + \cdots + \beta_{1k} F_{ik} + \epsilon_{i1} \\
X_{i2} - \mu_2 &= \beta_{21} F_{i1} + \cdots + \beta_{2k} F_{ik} + \epsilon_{i2} \\
&\vdots \\
X_{ip} - \mu_p &= \beta_{p1} F_{i1} + \cdots + \beta_{pk} F_{ik} + \epsilon_{ip}
\end{aligned}$$

Bonus: Factor Analysis

If there are fewer factors F_{ik} than features X_{ij}, then we have a reduced representation of the information in the features.

For all j, $X_{ij} - \mu_j = \sum_{k=1}^{q} \beta_{jk} F_{ik} + \epsilon_{ij}$ can be written as

$$\boldsymbol{X}_i - \boldsymbol{\mu} = \boldsymbol{\beta}\boldsymbol{F} + \boldsymbol{\epsilon}$$

Given X_{ij}'s, what are β_{jk}, F_{ik}, and ϵ_{ij}?

$$\begin{aligned} Cov(\boldsymbol{X}) &= Cov(\boldsymbol{\beta}\boldsymbol{F} + \boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta}\boldsymbol{\beta}^T + Cov(\boldsymbol{\epsilon}) \end{aligned}$$

and the off diagonal entries of $Cov(\boldsymbol{\epsilon})$ should be zero, so we look for factor loadings $\boldsymbol{\beta}$ to minimize the off diagonal entries of $Cov(\boldsymbol{\epsilon})$.

Given $\boldsymbol{\beta}$, $\boldsymbol{F}$ and $\boldsymbol{\epsilon}$ can be solved for. The $\boldsymbol{\beta}$ can be adjusted to account for rotations in $\boldsymbol{F}$ that allow for more interpretability.

Bonus: Independent Component Analysis (ICA)

ICA starts out the same way as factor analysis, i.e.,

$$X_{ij} - \mu_j = \sum_{k=1}^{q} \beta_{jk} F_{ik}$$

but often omits an error term ϵ as it's an effective procedure w/o it

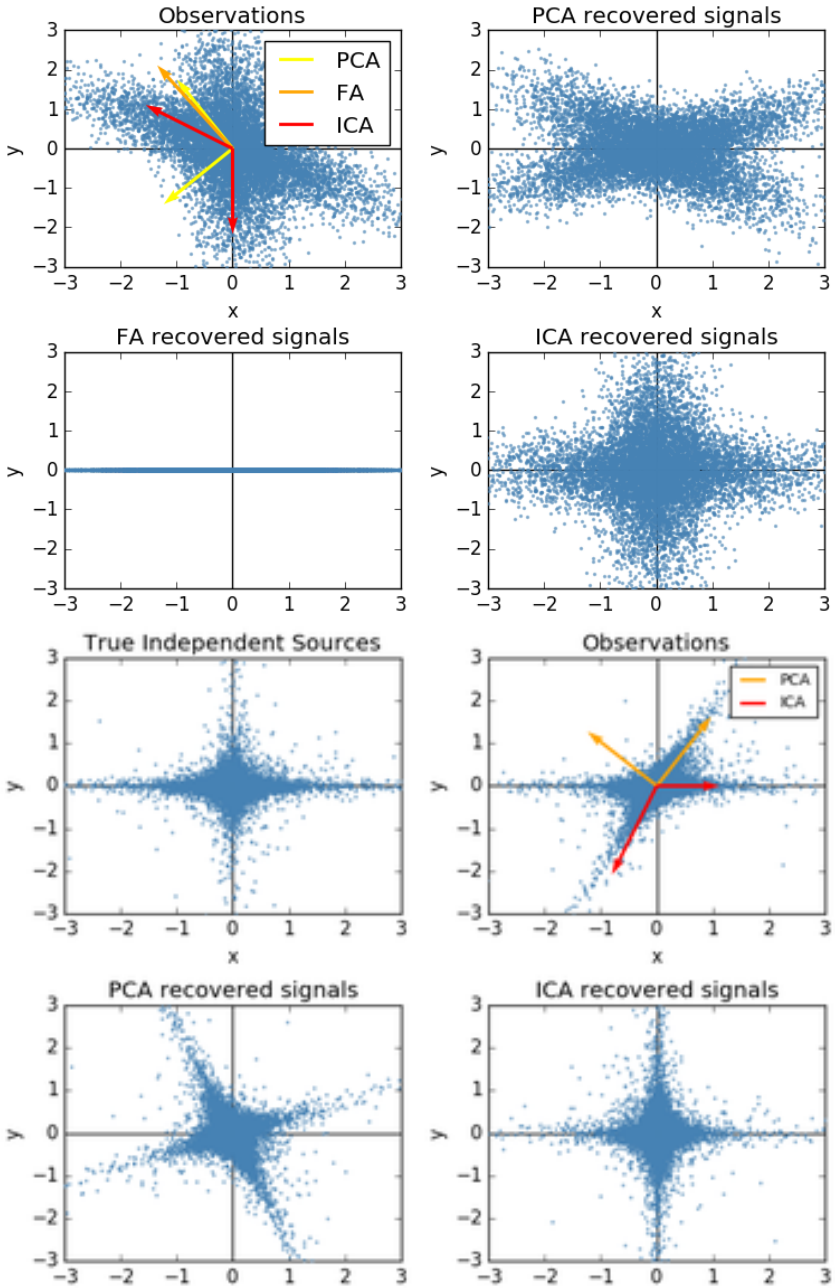
Only independence of F_{ik}'s is assumed – not marginal normality; in fact, the components are chosen sequentially based on *extremeness from normality* (as characterized by kurtosis, their 4^{th}-moment*).

ICA imagines each F_{ik} gives out independent signals picked up to varying degrees by different X_{ij} (i.e., *the cocktail party problem*).

---

*1^{st}-moment is mean, 2^{nd}-moment is variance, 3^{rd}-moment is skew. Kurtosis is a measure of "tail thickness," i.e., the extremeness of points that can be generated by the distribution in question.

Bonus: Graphically



Factor analysis finds an underlying uncorrelated lower dimension multivariate normal distribution that, *plus noise*, generates the observed higher dimension data cloud. E.g., 3D data generated from a 2D uncorrelated multivariate normal distribution:

$$X_{i1} = 1 \cdot F_{i1} + 0 \cdot F_{i2} + \epsilon_{i1}$$
$$X_{i2} = 1 \cdot F_{i1} + 0 \cdot F_{i2} + \epsilon_{i2}$$
$$X_{i3} = 0 \cdot F_{i1} + 1 \cdot F_{i2} + \epsilon_{i3}$$

ICA finds a basis representation of the data cloud such that the coordinates of the representation are independent of each other.

Bonus: PCA vs. Factor Analysis vs. ICA

The eigenvectors of the feature covariance matrix are an orthogonal basis whose axes capture the directions of maximum variation in the data. The eigenvalues are the relative amounts of variation captured along these axes. Transformation of the data to this coordinate system (i.e., *principal components*) is called PCA.

PCA is not a generative model like FA – PCA is purely a representative tool used to describe the data; nonetheless, both methods find orthogonal basis data representations; moreover, the (reduced representation) bases reconstruct the covariance matrix (PCA) or its off-diagonal entries (FA) so for a sufficient number, q, of factors and principal components and a relatively uninfluential diagonal, PCA and FA give similar representations. From here FA latent factors can be "rotated" for interpretation. E.g., a latent factor could be lined up with with an ICA component rather than a PCA variance direction. Unlike ICA, however, the collective set of factor analysis factors remain orthogonal.

In ICA the underlying basis *is not* orthogonal. This is because rather than targeting an *uncorrelated* representation like FA and PCA, ICA targets an *independent* representation. Interestingly, the detection of such a representation requires *exactly* that the underlying representations *not be normal* – the opposite of the FA assumption. Independent signals need not be orthogonal features... and interestingly these are found in ICA by simply maximizing the fourth (kurtosis) rather then the second (variance) moment.

Some Scratch Work

$$
\begin{aligned}
Var(X_1) &= 3 \\
Var(X_2) &= 3 \\
Cov(X_1, X_2) &= 2
\end{aligned}
$$

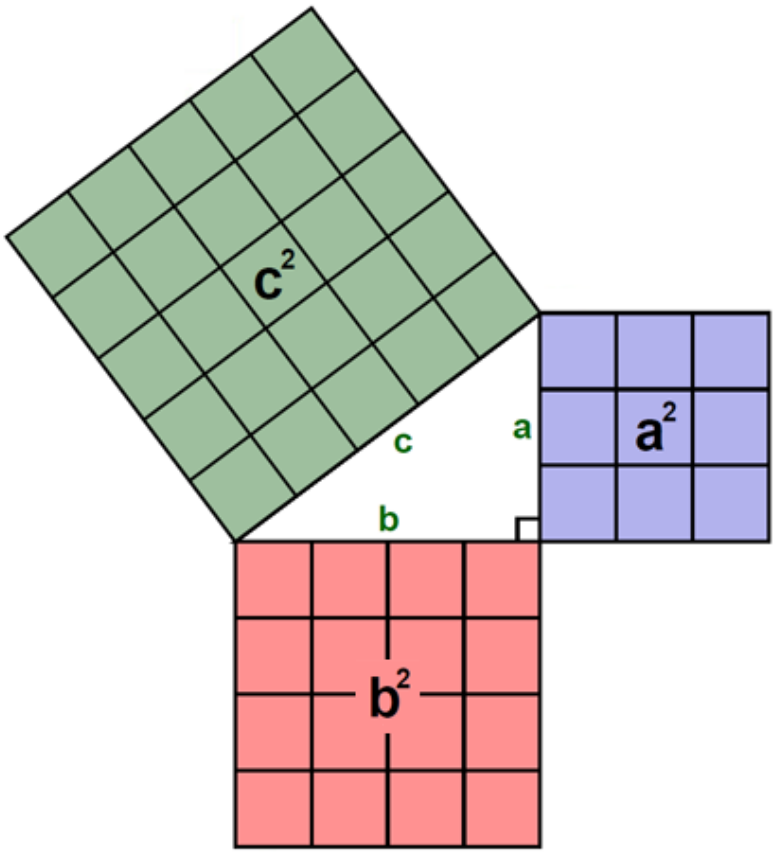
$$
\begin{aligned}
Var\left(\frac{1}{\sqrt{2}}X_1 + \frac{1}{\sqrt{2}}X_2\right) &= \left(\frac{1}{\sqrt{2}}\right)^2 Var(X_1) + \left(\frac{1}{\sqrt{2}}\right)^2 Var(X_2) \\
&\quad + 2\frac{1}{\sqrt{2}}\frac{1}{\sqrt{2}}Cov(X_1, X_2)
\end{aligned}
$$

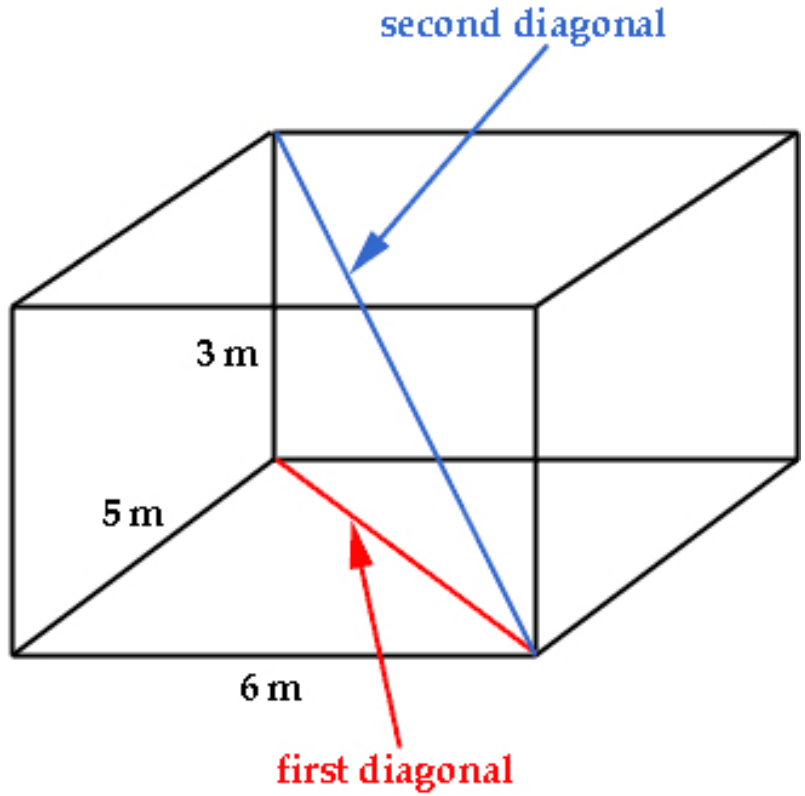
$$
\begin{aligned}
Var\left(\frac{1}{\sqrt{2}}X_1 + \frac{-1}{\sqrt{2}}X_2\right) &= \left(\frac{1}{\sqrt{2}}\right)^2 Var(X_1) + \left(\frac{-1}{\sqrt{2}}\right)^2 Var(X_2) \\
&\quad + 2\frac{1}{\sqrt{2}}\frac{-1}{\sqrt{2}}Cov(X_1, X_2)
\end{aligned}
$$

'Directional', Non P.D., & Singular

$$\begin{bmatrix} 1 & .6 & .4 \\ .6 & 1 & .4 \\ .4 & .4 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & .5 \\ 1 & .5 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 1 & .5 \\ 1 & 1 & .5 \\ .5 & .5 & 1 \end{bmatrix}$$

PCA as an *approximate* Pythagorean Theorem





$$c^2 = 5^2 + 6^2$$

$$c^2 = c^2 + 3^2$$

$$c^2 = 5^2 + 6^2 + 3^2$$