(PCA) PRINCIPLE COMPONENT ANALYSIS (HOTELLING 1933, PEARSON 1901)

SETTING: GIVEN A N (# OBSERVATIONS) BY d (# VARIABLE / SUBJECT) DATA MATRIX $\mathbb{X}$

$$\mathbb{X} = \begin{bmatrix} \vec{x}_1^{\,T} \\ \vdots \\ \vec{x}_N^{\,T} \end{bmatrix}$$

WE ASSUME $\vec{x}_1, \dots, \vec{x}_N$ ARE N I.I.D. RANDOM VECTORS IN $\mathbb{R}^d$ DRAWN FROM A DISTRIBUTION $F(\vec{x})$ W/ (UNKNOWN) MEAN μ AND COVARIANCE Σ

FIRST PRINCIPAL COMPONENT

ASSUME d IS LARGE. TO SAVE SPACE, FOR EACH $\vec{x}_i$ WE WANT A SCALAR y_i THAT CAPTURES AS MUCH OF THE VARIABILITY IN $\vec{x}_1, \dots \vec{x}_N$ AS POSSIBLE.

ASSUMPTION 1: y_i IS A LINEAR COMBINATION OF $\vec{x}_i$

$$y_i = \sum_{j=1}^{d} x_{ij} w_j = x_i^T w \quad , \; w \in \mathbb{R}^d$$

↖ SAME FOR ALL $\vec{x}_i$

NOTE:

ASSUMPTION 2: SAMPLE MEAN OF $\vec{x}_1, \dots \vec{x}_N$ IS $\vec{0}$; i.e. $\sum_{i=1}^{N} x_{ij} = 0$ FOR $j = 1, \dots, d$

NOTE: FOR ANY CHOICE OF $\vec{w}$

1) $\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} x_{ij} w_j = \frac{1}{N} \sum_{j=1}^{d} w_j \sum_{i=1}^{N} x_{ij} = 0$

$\Rightarrow$ SAMPLE MEAN OF $y_1, \dots y_N$ IS 0

2) IN PRACTICE, WE CENTER $\vec{x}_1, \dots \vec{x}_N$; $\vec{x}_i \longmapsto \vec{x}_i - \bar{x}$

GOAL: CHOOSE $\vec{w}$ TO MAXIMIZE SAMPLE VARIANCE OF $y_1, \ldots, y_N$

$$* \quad \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2 = \frac{1}{N}\sum_{i=1}^{N} y_i^2 = \frac{1}{N}\sum_{i=1}^{N}(\vec{x}_i^T\vec{\omega})^2 = \vec{\omega}^T\left(\frac{\sum_{i=1}^{N} x_i x_i^T}{N}\right)\vec{\omega}$$

$$\frac{1}{N}\sum_{i=1}^{N} y_i^2 = \frac{1}{N}\|y\|^2, \qquad \frac{1}{N}\sum_{i=1}^{N}(\vec{x}_i^T\vec{\omega})^2 = \frac{1}{N}\|X\omega\|^2, \qquad \vec{\omega}^T\left(\frac{\sum_{i=1}^{N} x_i x_i^T}{N}\right)\vec{\omega} = \vec{\omega}^T\left(\frac{X^TX}{N}\right)\vec{\omega}$$

NOTE: 1) WE CAN REPLACE $\vec{w}$ W/ $2\vec{w}$ AND INCREASE $*$ SO WE RESTRICT TO $\vec{w}$ S.T. $\|\vec{\omega}\|^2 = 1$

2)
$$\vec{w}_1 = \underset{\|\vec{\omega}\|^2=1}{\operatorname{argmax}} \; \vec{\omega}^T\left(\frac{X^TX}{N}\right)\vec{w}$$

GEOMETRIC INTERPRETATION OF $\vec{w}_1$

1) $\vec{w}_1$ IS UNIT LENGTH (SPECIFIES A DIRECTION IN $\mathbb{R}^d$)

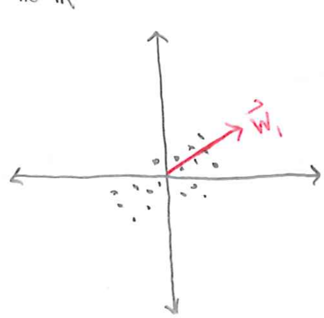
2) FOR ANY $\vec{x}_i$ WE HAVE

1)
$$\vec{x}_i = \underbrace{\left(\vec{x}_i - (x_i^T\omega_1)\omega_1\right)}_{x_{i\perp}} + \underbrace{(x_i^T\omega_1)\,\omega_1}_{x_{i\parallel}}$$

2)
$$\|\vec{x}_i\|^2 = \|\vec{x}_i - (x_i^T\vec{\omega}_1)\vec{\omega}_1\|^2 + \underbrace{(x_i^T\vec{w}_1)^2}_{y_i^2}$$

$\Rightarrow$ FINDING $\vec{\omega}_1$ THAT MAXIMIZES SAMPLE VARIANCE OF $y_1, \ldots y_N$ IS EQUIVALENT TO FINDING DIRECTION ALONG WHICH $\vec{x}_1, \ldots, \vec{x}_N$ MOST GREATLY VARIES

EX: IN $\mathbb{R}^2$



CALCULATING $\vec{w}_1$

- TO FIND $\vec{w}_1$ WE WANT TO MAXIMIZE

$$f(\vec{w}) = \vec{w}^T \left(\frac{X^T X}{N}\right) \vec{w}$$

SUBJECT TO CONSTRAINT $\|\vec{w}\|^2 = 1$ } LAGRANGE MULTIPLIERS!

- LET $g(w,\lambda) = f(\vec{w}) + \lambda\left(\|w\|^2 - 1\right)$ AND $\nabla = \left(\frac{\partial}{\partial w_1}, \ldots, \frac{\partial}{w_d}\right)$

$$\vec{0} = \nabla g(\vec{w},\lambda) = \nabla\left(\vec{w}^T \frac{X^T X}{N} \vec{w}\right) - 2\lambda\vec{w} = 2\left(\frac{X^T X}{N}\vec{w} - \lambda\vec{w}\right)$$

$$1 = \|w\|^2$$

$$\Rightarrow \left(\frac{X^T X}{N}\right)\vec{w} = \lambda\vec{w} = \hat{\Sigma}\,\vec{w}$$

$$\|w\|^2 = 1$$

RECALL, $\hat{\Sigma} = \frac{1}{N} X^T X - \frac{1}{N^2} X^T 1_N \underbrace{1_N^T X}_{=\vec{0}}$

$$= \frac{1}{N} X^T X$$

- THERE ARE (UP TO) d SOLUTIONS TO $\hat{\Sigma}\vec{w} = \lambda\vec{w}$.

  * PICK ONE ASSOCIATED W/ LARGEST EIGENVALUE!

$$\Rightarrow f(\vec{w}) = \vec{w}_1^T \hat{\Sigma} w_1 = w_1^T (\lambda_{MAX}) \vec{w}_1$$
$$= \hat{\lambda}_{MAX}$$

* THERE IS NOT A UNIQUE SOLUTION
1) $-\vec{w}_1$ ALSO WORKS
2) λ_{MAX} MAY HAVE 2+ DIMENSIONAL EIGENSPACE!

* MAKE A CHOICE!

- CALL $y_i = \vec{x}_i^T \vec{w}_1$ THE FIRST PRINCIPAL COMPONENT SCORE OF $\vec{x}_i$

ADDITIONAL PRINCIPAL COMPONENTS

OUR CHOICE $\vec{w}_1$ HAS FOUND THE/A DIRECTION OF GREATEST VARIABILITY IN OUR DATA, BUT THE DATA MAY STILL VARY GREATLY IN OTHER DIRECTIONS

IDEA: 1) FOR EACH x_i REMOVE PORTION $\parallel$ TO $\vec{w}_1$

$$\vec{x}_i^{(1)} = \vec{x}_i - (x_i^T \vec{w}_1) w_1$$

2) REPEAT TO GET $\vec{w}_2$ = DIRECTION OF 2^{ND} GREATEST VARIABILITY

$$\vec{x}_i^{(2)} = \vec{x}_i^{(1)} - (\vec{x}_i^{(1)} \vec{w}_2) w_2$$

3) AND SO ON

NOTES:

1) IN MATRIX NOTATION

$$X^{(k)} = \begin{bmatrix} \vec{x}_1^{(k)T} \\ \vdots \\ \vec{x}_N^{(k)T} \end{bmatrix} = \begin{bmatrix} \vec{x}_1^T - \sum_{s=1}^{k} (\vec{x}_1^T \vec{\omega}_s) \omega_s^T \\ \vdots \\ \vec{x}_N^{(k)T} - \sum_{s=1}^{k} (\vec{x}_N^T \vec{\omega}_s) \vec{\omega}_s \end{bmatrix} = X - \sum_{s=1}^{k} X \vec{\omega}_s \omega_s^T$$

2) THE k PRINCIPAL COMPONENT SCORES OF $\vec{x}_i$ ARE

$$\vec{y}_i^{\,T} = (\vec{x}_i^T \vec{\omega}_1, \ \vec{x}_i^T \vec{w}_2, \ \ldots \ \vec{x}_i^T \vec{w}_k) = x_i^T (\vec{w}_1, \ldots \vec{w}_k)$$

$\vec{w}_i$ IS CALLED THE ith VECTOR OF PRINCIPAL COMPONENT LOADINGS.

3) $\vec{W}_S$ AND $\vec{W}_T$ ARE PERPENDICULAR FOR $S \neq T$

$\underline{S > T}$: $X^{(S-1)}$ IS COMPRISED OF DATA W/ COMPONENTS PARALLEL TO $\vec{W}_T$ REMOVED

$\Rightarrow X^{(S-1)}$ DOES NOT VARY IN $\vec{W}_T$ DIRECTION AT ALL!

4) SINCE $\bar{x} = 0$ (BY ASSUMPTION EARLIER)

a) $\bar{y} = \frac{1}{N} \sum_{i=1}^{N} \vec{y}_i = \vec{0}$

b) $\hat{\Sigma}_y = \frac{1}{N} \sum_{i=1}^{N} (\vec{y}_i - \bar{y})(\vec{y}_i - \bar{y})^T = \frac{1}{N} \sum_{i=1}^{N} \vec{y}_i \vec{y}^T = \frac{1}{N} Y^T Y$

$\in \mathbb{R}^{k \times k}$

$$Y = \begin{bmatrix} \vec{y}_1^{\,T} \\ \vdots \\ \vec{y}_N^{\,T} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{N} \sum_{i=1}^{N} y_{i1}^2 & & 0 \\ & \ddots & \\ 0 & & \frac{1}{N} \sum_{i=1}^{N} y_{ik}^2 \end{bmatrix}$$

$\frac{1}{N} \sum_{i=1}^{N} y_{is} y_{it} = \frac{1}{N} \sum_{i=1}^{N} (\vec{x}_i \vec{w}_s^T) \underbrace{(\vec{w}_t \vec{x}_i)}_{=0}$

$= 0$

IN-CLASS EXERCISE

$$= \begin{bmatrix} \hat{\lambda}_1 & & 0 \\ & \ddots & \\ 0 & & \hat{\lambda}_k \end{bmatrix} = \Lambda_k$$

$*$ $\vec{x}_i \in \mathbb{R}^d \xrightarrow{PCA} \vec{y}_i \in \mathbb{R}^k$

UNCORRELATED COMPONENTS

Recap of PCA

Given data matrix $X \in \mathbb{R}^{N\times d}$ (w/ mean zero columns) and sample covariance matrix $\hat{\Sigma} = \dfrac{X^T X}{N} = \sum_{i=1}^{N} \vec{x}_i \vec{x}_i^T / N$

1) $\hat{\Sigma} \xrightarrow{\text{PCA}}$ First principal loading $\vec{w}_1$ $\longrightarrow$ First prin. scores $y_{i1} = \vec{x}_i^T \vec{w}_1$

" " variance λ_1

$$\hat{\Sigma} \vec{w}_1 = \lambda_1 \vec{w}_1$$

2) $X^{(1)} = X - (X^T \vec{w}_1) w_1^T \longrightarrow \hat{\Sigma}^{(1)} = \dfrac{X^{(1)T} X}{N}$

(Remove component of $\vec{x}_i$ // to $\vec{w}_1$)

$\downarrow$ PCA

2^{nd} principal loading $\vec{w}_2$

" " variance λ_2

$$\hat{\Sigma}^{(1)} \vec{w}_2 = \lambda_2 \vec{w}_1$$

$\downarrow$

2^{nd} principal score

$$y_{i2} = \vec{x}_i^T \vec{w}_2$$

3) $X^{(2)} = X^{(1)} - \left(X^{(1)} \vec{w}_2\right) \vec{w}_2^T$ and continue ...

$\vdots$

Principal component vector $\vec{w}_1, \dots, \vec{w}_d$

" " variances $\lambda_1 \geq \lambda_2, \dots \geq \lambda_d$

scores $\vec{y}_i = \begin{bmatrix} \vec{w}_1^T \vec{x}_i \\ \vdots \\ \vec{w}_d^T \vec{x}_i \end{bmatrix} = \begin{bmatrix} \vec{w}_1^T \\ \vdots \\ \vec{w}_d^T \end{bmatrix} \vec{x}_i$

CLAIM: SUPPOSE $\vec{v}_1, \ldots, \vec{v}_d$ ARE LINEARLY IND. EIGENVECTOR OF $\hat{\Sigma} = \frac{X^T X}{N}$

W/ EIGENVALUES $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$ RESPECTIVELY

$$\left(\begin{array}{l} \underline{\text{NOTE}}: \hat{\Sigma} \text{ IS SYMMETRIC} \Rightarrow \lambda_1 \ldots \lambda_d \in \mathbb{R}^d \\ \qquad\qquad\qquad\qquad\quad \Rightarrow \vec{w}_1 \ldots \vec{w}_d \text{ ARE MUTUALLY ORTHOGONAL} \end{array} \right)$$

THEN $\vec{w}_1, \ldots \vec{w}_d$ ARE EIGENVECTORS OF $\hat{\Sigma}^{(1)} = \frac{X^{(1)T} X^{(1)}}{N}$

W/ EIGENVALUES $0, \lambda_2, \ldots, \lambda_d$ RESPECTIVELY

WHY?

ONE CAN SHOW (HOMEWORK)

$$\hat{\Sigma}^{(1)} = \hat{\Sigma} - \lambda_1 \vec{w}_1 \vec{w}_1^T$$

$$\begin{cases} \hat{\Sigma}^{(1)} \vec{w}_1 = \left(\hat{\Sigma} - \lambda \vec{w}_1 \vec{w}_1^T \right) \vec{w}_1 = \lambda \vec{w}_1 - \lambda \vec{w}_1 = 0 \\ \hat{\Sigma}^{(1)} \vec{w}_j = \left(\hat{\Sigma} - \lambda \vec{w}_1 \vec{w}_1^T \right) \vec{w}_j = \lambda_j \vec{w}_1 \qquad j \neq i \end{cases}$$

GENERALIZING WE COULD SIMPLY COMPUTE ALL EIGENVECTOR / EIGENVALUES FROM $\hat{\Sigma}$ TO START.

$$\hat{\Sigma} \xrightarrow{\text{SPECTRAL DECOMPOSITION}} \begin{array}{l} \text{PRINCIPAL LOADINGS } \vec{w}_1, \ldots \vec{w}_d \\ \text{PRINCIPAL VARIANCES } \lambda_1 \geq \lambda_2, \ldots \geq \lambda_d \end{array}$$

$$\downarrow$$

SCORES $\vec{y}_i$

Spectral Decomposition & PCA

- Letting $W = [\vec{w}_1 | \ldots | \vec{w}_d]$ and $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$ be eigenvalues of $\hat{\Sigma}_X$

$$\hat{\Sigma}_X \vec{W} = [\lambda_1 \vec{w}_1 | \ldots | \lambda_d \vec{w}_d] = \vec{W} \Lambda \qquad \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{bmatrix}$$

Since columns of $\vec{W}$ are unit length and orthogonal $\vec{W}$ is orthonormal

$$\vec{\Sigma}_X = \vec{W} \Lambda W^T \qquad , \vec{W} W^T = I$$

- Observe

$$\vec{y}_i^T = (\vec{x}_i^T \vec{w}_1, \ldots, \vec{x}_i \vec{w}_d) = \vec{x}_i^T W$$

so that

$$Y = \begin{bmatrix} \vec{y}_1^T \\ \vdots \\ \vec{y}_N \end{bmatrix} = X W$$

matrix of principle scores

- By construction means of columns of $Y = 0$ (since $\bar{x} = 0$)

$$\Rightarrow \hat{\Sigma}_Y = \frac{Y^T Y}{N} = \frac{W^T X^T X W}{N} = W^T \hat{\Sigma}_X W$$
$$= \Lambda$$

1) Components of principal scores are uncorrelated !

2) Variance of principal scores are decreasing !

3) $\underset{\text{"total variation" in } X}{tr(\hat{\Sigma}_X)} = tr(\Lambda) = \underset{\text{"total variation" in } Y}{tr(\hat{\Sigma}_Y)}$

Geometric Interpretation of PCA

- Given data in $\mathbb{R}^d$, PCA

  1) learns an orthogonal basis $\{\vec{w}_1, \ldots, \vec{w}_d\}$ of $\mathbb{R}^d$ — loadings

  2) learns coordinates of each datum in this basis

     "Coordinates of $\vec{x}_i$ in basis $\{\vec{w}_1, \ldots, \vec{w}_d\}$ are $\vec{y}_i$" — scores of $\vec{x}_i$

such that <u>in this basis</u>

  1) the coordinates are uncorrelated

  2) given in decreasing order of variance

$$\begin{aligned} \text{variance of } 1^{\text{st}} \text{ coordinate} &= \lambda_1 \\ &\geq \\ \text{" " } 2^{\text{nd}} \text{ "} &= \lambda_2 \\ &\vdots \\ \text{" " } d^{\text{th}} \text{ "} &= \lambda_d \end{aligned}$$

- We can approximate each datum using only the first k principal components

$$\vec{x}_i^{(k)} = y_{i1}\vec{w}_1 + \ldots + y_{ik}\vec{w}_k = \left[\vec{w}_1 \mid \ldots \mid \vec{w}_k\right] \begin{bmatrix} y_{i1} \\ \vdots \\ y_{ik} \end{bmatrix}$$

  1) $\left.\begin{array}{l} kd \text{ values in loadings } \vec{w}_1, \ldots \vec{w}_k \\ Nk \text{ " in scores } (y_{i1}, \ldots y_{ik}) \\ \qquad\qquad i = 1, \ldots N \end{array}\right\}$ $k(N+d)$ values in prin. comp

     (Nd values in original data)

  2) $\vec{x}_1^{(k)}, \ldots, x_N^{(k)}$ are elements in span $\{\vec{w}_1, \ldots, \vec{w}_k\}$

     $\therefore$ approximating original data w/ approximates on a k-dimensional subspace!

THM: GIVEN ANY k-DIMENSIONAL LINEAR SUBSPACE V OF $\mathbb{R}^d$, LET

$\text{PROJ}_V(\vec{x}_i)$ BE THE ORTHOGONAL PROJECTION OF $\vec{x}_i$ ONTO V.

THEN

$$\frac{1}{N}\sum_{i=1}^{N} \| \vec{x}_i - \text{PROJ}_V(\vec{x}_i) \|^2 \geq \lambda_{k+1} + \dots + \lambda_d$$

IF $\lambda_k > \lambda_{k+1}$ THEN $\text{SPAN}\{\vec{w}_1, \dots \vec{w}_k\}$ IS THE UNIQUE k-DIMENSIONAL LINEAR SUBSPACE FOR WHICH

$$\frac{1}{N}\sum_{i=1}^{N} \| \vec{x}_i - \text{PROJ}_V(\vec{x}_i) \|^2 = \lambda_{k+1} + \dots + \lambda_d$$

" PCA LEARNS THE k-DIMENSIONAL LINEAR SUBSPACE WHICH IS CLOSEST TO DATA ON AVERAGE"

COROLLARY:

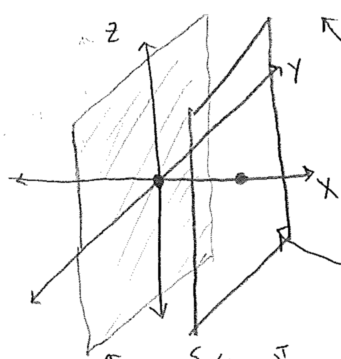
REMOVE THE $\bar{x} = 0$ ASSUMPTION. GIVEN ANY k DIMENSIONAL AFFINE SUBSPACE V OF $\mathbb{R}^d$

$$\frac{1}{N}\sum_{i=1}^{N} \| \vec{x}_i - \text{PROJ}_V(\vec{x}_i) \|^2 \geq \lambda_{k+1} + \dots + \lambda_d$$

AND $\lambda_k > \lambda_{k+1}$ IMPLIES $\bar{x} + \text{SPAN}\{\vec{w}_1, \dots \vec{w}_k\}$ IS THE UNIQUE k-DIMENSIONAL AFFINE SUBSPACE ATTAINING THE MINIMUM

NOTE: $\bar{x} + \text{SPAN}\{\vec{w}_1, \dots, \vec{w}_k\} = \{\vec{v} \mid \vec{v} - \bar{x} \in \text{SPAN}\{\vec{w}_1, \dots \vec{w}_k\}\}$

AFFINE SUBSPACE = SHIFTED LINEAR SUBSPACE



$\{(1,0,0)^T\} + \text{SPAN}\{(1,0,0)^T, (0,1,0)^T\}$

$\text{SPAN}\{(1,0,0)^T, (0,1,0)^T\}$

2) IN PRACTICE

$$\vec{x}^{(k)} = \bar{x} + \sum_{s=1}^{k} y_{ik} \vec{w}_k$$

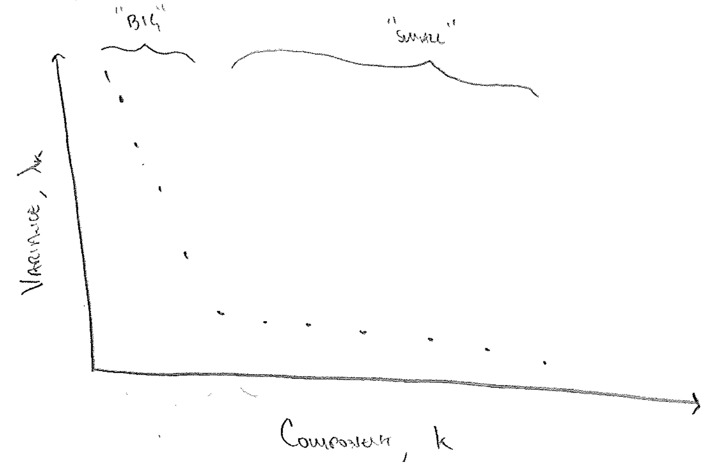
REMOVE MEAN → COMPUTE PCA

→ FIND APPROX.

→ ADD MEAN BACK

ASSESSING PERFORMANCE OF PCA

- SCREE PLOT



IDEALLY WE WOULD TO OBSERVE A CLEAR SEPARATION OF MAGNITUDE OF VARIANCES / EIGENVALUES

1) NO GUARANTEE THOUGH EXPONENTIAL / RAPID DECAY IS OBSERVED IN MANY SETTINGS

2) NO AGREED UPON RULES

- CONSIDER PROPORTION OF VARIANCE EXPLAINED BY FIRST λ PRIN. COMP.

$$\sum_{s=1}^{k} \lambda_s \Big/ \sum_{s=1}^{d} \lambda_s = \frac{\sum_{s=1}^{k} \lambda_s}{tr(\hat{\Sigma})} \in [0,1]$$

DIFFERENT AUTHORS PROPOSE 80%, 90%, ETC. FOR DETERMINING CUTOFF k

3) $\lambda_s = 0$ FOR $s > k$ IMPLIES DATA RESIDES ON k-DIMENSIONAL HYPERPLANE

RECALL PCA BREAKS VARIABILITY IN DATA INTO VARIABILITY IN ORTHOGONAL DIRECTIONS

1) LEARN NEW BASIS OF $\mathbb{R}^d$, $\vec{w}_1, \ldots, \vec{w}_d$

2) BASIS SORTED BY $\mathrm{Var}(\vec{w}_i^T \vec{x}) = \lambda_i$

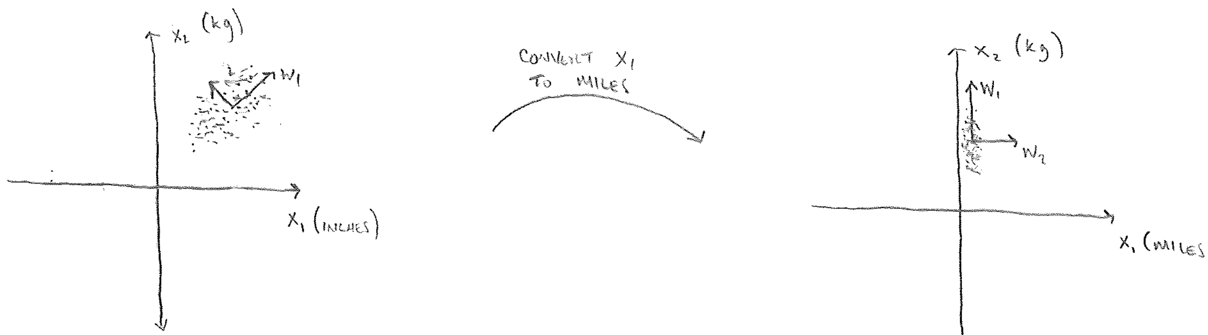
$\Rightarrow$ RELIANCE ON VARIANCE & LINEARITY MAKES PCA

1) SENSITIVE TO SCALE IN DATA (INCLUDING OUTLIERS)

2) STRUGGLE W/ THE DETECTION ON NONLINEAR/AFFINE SUBSPACES

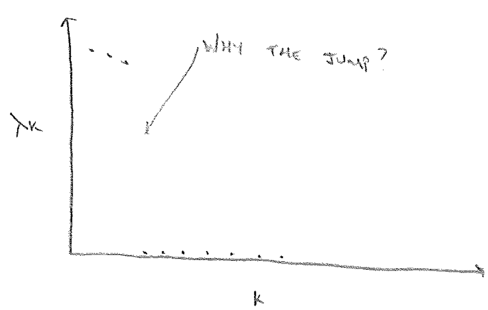
SCALE:

Ex:



$\lambda_1 \approx \lambda_2 > 0$ $\qquad\qquad$ $\lambda_1 \gg \lambda_2 \approx 0$

QUESTION: LOW DIMENSIONALITY OF DATA OR BIAS ARISING FROM IMBALANCED SCALES

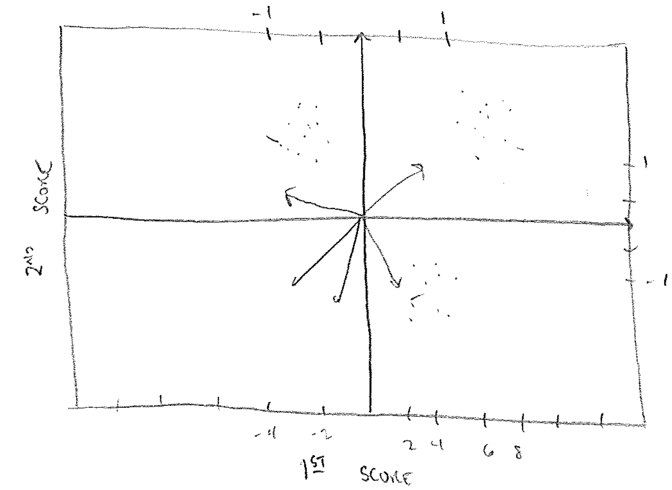


OPTION 1: STANDARDIZE THE DATA

$$x_{ij} \longmapsto \frac{x_{ij} - \bar{x}_j}{\sigma_j} \qquad \bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}, \quad \sigma_j^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2$$

* THIS IS EQUIVALENT TO FINDING EIGENDECOMPOSITION OF SAMPLE CORRELATION MATRIX!

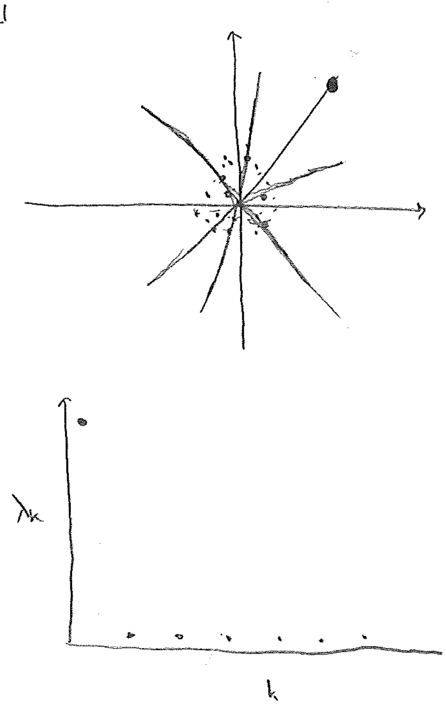
OPTION 2: USE A BIPLOT TO ASSESS INFLUENCE OF POINTS & INDIVIDUAL COVARIATES



- DOTS ARE 1ST/2ND SCORES OF DATA (y_{i1}, y_{i2})
  - SCALE SHOWN ON LEFT/LOWER AXES
- VECTORS ARE COMPONENTS OF 1ST/2ND LOADINGS (w_{1j}, w_{2j})
  - SCALE SHOWN ON UPPER/RIGHT AXES

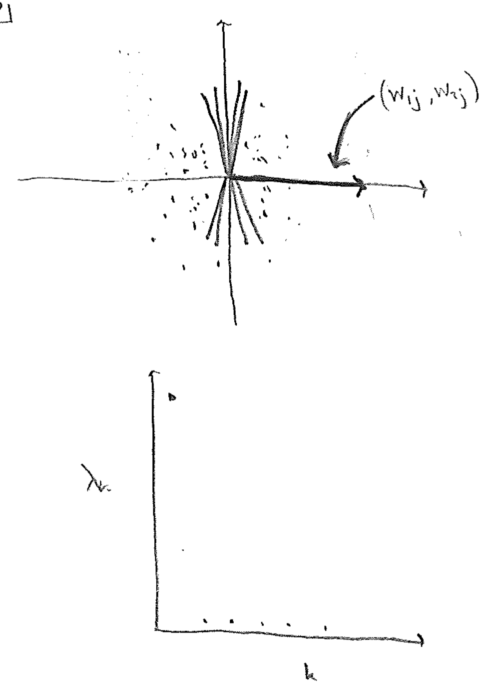
INTERPRET THE FOLLOWING BIPLOT SCREE PLOTS PAIRS

A)



FIRST LOADING IS DETERMINED BY OUTLIER IN DATA

B)

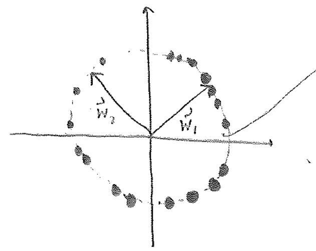


FIRST LOADING IS DETERMINED BY j^{TH} COVARIATE!

* RESCALING DATA

LINEARITY:

Ex: DATA IN $\mathbb{R}^2$ DRAWN UNIFORMLY FROM UNIT CIRCLE



DATA IS 1-DIMENSIONAL

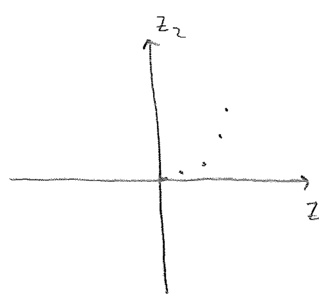
ON AVERAGE $\lambda_1 \approx \lambda_2 \approx 1 \implies$ FROM PERSPECTIVE OF PCA DATA RESIDES IN $\mathbb{R}^2$

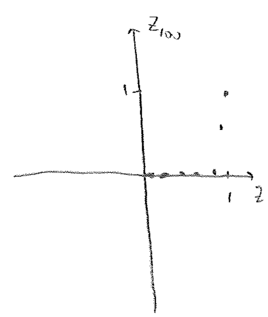
* NO LOWER DIMENSIONAL GEOMETRY DETECTED!

Ex: CONSIDER $z_1, \dots z_N \sim U(0,1)$. LET $\vec{x}_i = \begin{bmatrix} z_i \\ z_i^2 \\ \vdots \\ z_i^d \end{bmatrix} + \varepsilon_i$ * DATA IS 1-DIMENSIONAL

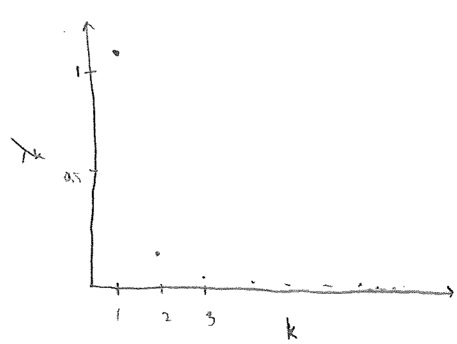
$\varepsilon_i \sim N(\vec{0}, 0.1 I_d)$

KNOW $x_1 \mapsto$ KNOW $x_2, \dots x_d$





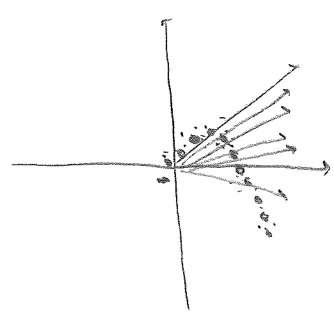
SCREE



$\lambda_s \approx 0 \quad s > 2$

SUGGESTS DATA IS $\approx$ 2D

BIPLOT



* SHOWS APPARENT FUNCTIONAL RELATIONSHIP

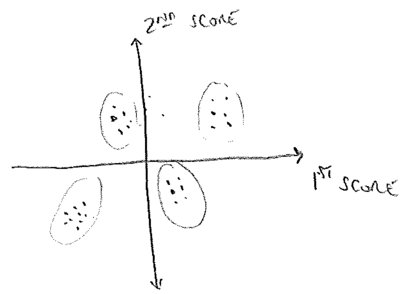
HINTS AT NONLINEARITY!

INTERPRETABILITY OF SCORES

$$\vec{y}_i = \begin{bmatrix} \vec{w}_1^T x_i \\ \vdots \\ \vec{w}_d^T \vec{x}_i \end{bmatrix} = \begin{bmatrix} \vec{w}_1^T \\ \vdots \\ \vec{w}_d^T \end{bmatrix} \vec{x}_i = \left(\sum_{j=1}^{d} x_{ij} \vec{w}_{1j}, \ldots, \sum_{j=1}^{d} x_{ij} w_{dj} \right)$$

* SCORES ARE A LINEAR COMBINATION OF d OBSERVED VARIABLES

1) IN SOME CASES, THE FIRST FEW SCORES CAN IDENTIFY GROUPS IN DATA



2) CAN USE SCORE $(y_{i1}, .. y_{ik})$ w/ $k << d$ IN MODELING

   i) SINCE $\vec{y}_i^{(k)} = (y_{i1}, .. y_{ik})$ IS IN MUCH LOWER DIMENSION THAN $\vec{x}_i$ <u>HOPE</u> WE CAN ESCAPE/RELAX THE CURSE OF DIMENSIONALITY

   ii) HOW DO WE INTERPRET $y_{ij}^{(k)}$? UNITS? PHYSICAL INTERPRETATION? CONNECTION TO OUTCOME OR FEATURES OF INTEREST?

<u>OBSERVE:</u> a) IF ALL COVARIATES HAVE SAME UNITS "L" THEN IT IS NATURAL TO SAY y_{ij} HAS UNITS "L".

b) IF COVARIATES HAVE MIXED UNITS WHAT ARE UNITS OF y_{ij}? HOW DOES THIS DEPEND ON $\vec{w}_j$?

$$\left.\begin{array}{l} \vec{x}_i^T = (L_1, L_1, L_2, L_1, L_2, L_2) \\ w_j = (\_, \_, 0, \_, 0, 0) \end{array}\right\} \; y_{ij} \text{ HAS UNITS } L_1 \, ?$$

WE CAN STANDARDIZE $\vec{x}_1, \ldots \vec{x}_n$ SO ALL COVARIATES ARE <u>UNITLESS</u> BUT OTHER ISSUES OF INTERPRETABILITY REMAIN!