Principal Component Analysis

Statistics 407
ISU

Definition

Principal component analysis is concerned with summarizing the _____ matrix. PCA involves computing the eigenvectors and eigenvalues of the variance-covariance matrix, S, or correlation matrix, R, as the first step. The _____ are used to _____ the data from p dimensions down to a lower dimensional representation. The _____ give the _____ of the data in the direction of the eigenvector. The first eigenvector is the vector which defines the direction of maximum variance in the data.

Objectives

The objectives of principal component analysis are

- *To _____. That is, if the data is plotted in p-dimensional space does it fill up all p dimensions. If not then the true dimensionality of the data may be less than p, and we can use a smaller number of variables to describe the data.*
- *_____ new meaningful underlying variables. The new variables are not always meaningful, but they may still be convenient.*

Common Uses

- *_____: finding outliers, finding strong association, finding clusters. Principal component analysis is badly affected by outliers, hence the practical usage for finding outliers.*
- *_____: Find a low-dimensional projection of the data which reveals clusters.*
- *Discriminant analysis, Regression analysis: Removing multicollinearity among explanatory variables.*

Definition

- *Principal components are _____ of the original variables. The _____ principal component is:*

$$\mathbf{Y}_1 = \mathbf{X}\mathbf{e}_1 = \begin{bmatrix} e_{11}X_{11} + e_{21}X_{12} + \ldots + e_{p1}X_{1p} \\ e_{11}X_{21} + e_{21}X_{22} + \ldots + e_{p1}X_{2p} \\ e_{11}X_{31} + e_{21}X_{32} + \ldots + e_{p1}X_{3p} \\ \vdots \\ e_{11}X_{n1} + e_{21}X_{n2} + \ldots + e_{p1}X_{np} \end{bmatrix} = \begin{bmatrix} Y_{11} \\ Y_{21} \\ Y_{31} \\ \vdots \\ Y_{n1} \end{bmatrix}$$

- *Where the projection, $\mathbf{e}_1$, is found by finding the _____ of maximum variance of the data in the multivariate space.*

Definition

- *The k'th principal component is:*

$$\mathbf{Y}_k = \mathbf{X}\mathbf{e}_k = \begin{bmatrix} e_{1k}X_{11} + e_{2k}X_{12} + \ldots + e_{pk}X_{1p} \\ e_{1k}X_{k2} + e_{2k}X_{22} + \ldots + e_{pk}X_{2p} \\ e_{1k}X_{31} + e_{2k}X_{32} + \ldots + e_{pk}X_{3p} \\ \vdots \\ e_{1k}X_{n1} + e_{2k}X_{n2} + \ldots + e_{pk}X_{np} \end{bmatrix} = \begin{bmatrix} Y_{11} \\ Y_{21} \\ Y_{31} \\ \vdots \\ Y_{n1} \end{bmatrix}$$

- *Where the projection, $\mathbf{e}_k$, is found by finding the **direction** of maximum variance of the data _____ projections in the multivariate space.*

Eigen-decomposition

- *Mathematically, the directions of maximum variation can be solved using an eigen-decomposition of the variance-covariance (or correlation) matrix.*

$$\mathbf{S} = \mathbf{E}'\Lambda\mathbf{E}$$

where $\mathbf{E} = \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1p} \\ e_{21} & e_{22} & \dots & e_{2p} \\ \vdots & \vdots & & \vdots \\ e_{p1} & e_{p2} & \dots & e_{pp} \end{bmatrix} = [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \dots \quad \mathbf{e}_p]$ is the matrix of eigenvectors, and $\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & & \\ \vdots & & \dots & \\ 0 & & & \lambda_p \end{bmatrix}$ is a diagonal matrix of eigenvalues.

The eigenvectors have length equal to 1 and are orthogonal ($\mathbf{e}_j'\mathbf{e}_k = 0, j \neq k$) to each other.

Principal Component Scores

- *The centered principal components are called the _____:*

$$\mathbf{Y}_k = (\mathbf{X} - \bar{\mathbf{X}})\mathbf{e}_k, k = 1, \dots p$$

- *These are the _____ that the cases take on the new set of variables.*

Total Variance

- *The _____ is the sum of the diagonal elements of the variance-covariance (or correlation) matrix:*

$$s_{11}+s_{22}+\ldots+s_{pp} = \sum_{i=1}^{p} Var(\mathbf{X}_i) = \lambda_1+\lambda_2+\ldots+\lambda_p = \sum_{i=1}^{p} Var(\mathbf{Y}_i).$$

- *It is equal to the ____ of the eigenvalues.*
- *Total variance is used to decide how many principal components to keep. The _____ of variation explained by the first k PCs is:* $\frac{\lambda_1+\ldots+\lambda_k}{\lambda_1+\ldots+\lambda_p}$

Properties of PCs

- *The _____ of the k'th principal component is equal to the value of the k'th eigenvalue,* λ_k .
- *Two principal components are _____.*
- *The total variance of the principal components is equal to the total variance of the original data, if all p principal components are used.*
- *The total variance is equal to ___ when the correlation matrix is used.*

Importance of Variables

- *The _____ between the i'th PC and the k'th variable is* $\frac{e_{ik}\sqrt{\lambda_i}}{s_k}, \quad i,k = 1,\ldots,p.$
- *This helps to determine how much variable k _____ to PC i.*

Example: Track records

- *This data contains the women's national records for 100m, 200m, 400m, 800m, 1500m, 3000m and marathon, for 55 countries. We will look at the first two track events, 100m and 200m, for all the countries.*

country	m100 (sec)	m200 (sec)	m400 (sec)	m800 (min)	m1500 (min)	m3000 (min)	marathon (min)
argentin(SA)	11.61	22.94	54.50	2.15	4.43	9.79	178.52
australi(PC)	11.20	22.35	51.08	1.98	4.13	9.08	152.37
austria(EU)	11.43	23.09	50.62	1.99	4.22	9.34	159.37
...							

Results for S

$$\mathbf{S} = \begin{bmatrix} 0.204 & 0.479 \\ 0.479 & 1.234 \end{bmatrix}$$

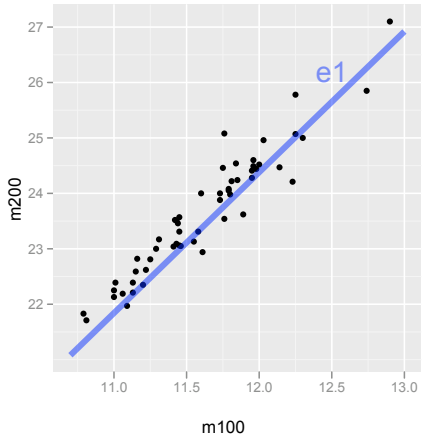
Variable	$\mathbf{e}_1$	$\mathbf{e}_2$
100m	0.366	0.931
200m	0.931	-0.366
Variance	1.423	0.016
% Tot Var	98.9	100.0

- *PC 1 explains _____ of the variation. This would suggest that _ PC would be enough to summarize the variation in this data.*
- *The correlation between PC 1 and the 100m is 0.97, and with the 200m is 1.00, so both variables are _____.*
- *The principal component scores for Argentina are:*

$$Y_{11} = 0.366 \times (11.61 - 11.62) + 0.931 \times (22.94 - 23.64) = -0.656,$$
$$Y_{12} = 0.931 \times (11.61 - 11.62) - 0.366 \times (22.94 - 23.64) = 0.266$$

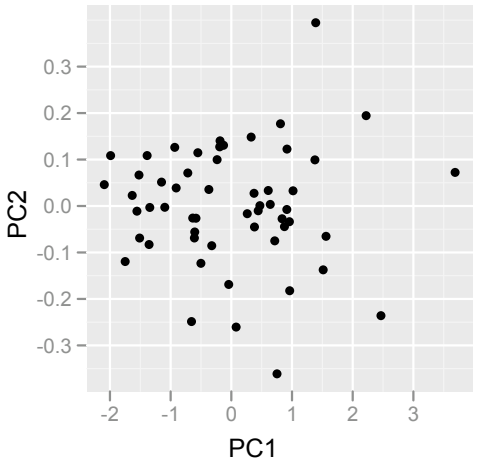
where the mean of 100m is 11.62, and for 200m is 23.64.

Results for S

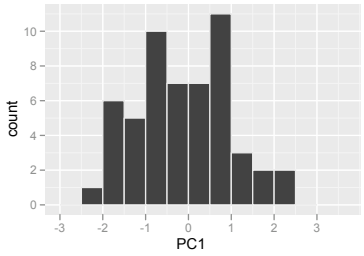


- *The first eigenvector is shown. This is the direction of _____ for this data. Its plotted here close to the data.*
- *First principal component scores are obtained by _____ the points onto this line.*
- *The _____ eigenvector is orthogonal to this one.*

Results for S



- *There should be _____ obvious structure to the new variables, PCs.*
- *PC 1 should be _____ to PC2.*



When to use R, not S?

- *PCA finds directions of maximum variance. If some variables have much _____ variance than the other variables, the PCA will simply return _____ large variance variables as the first few PCs.*
- *Use the _____, if there are big differences in the variances. This is the same as first _____ the variables.*

Results for R

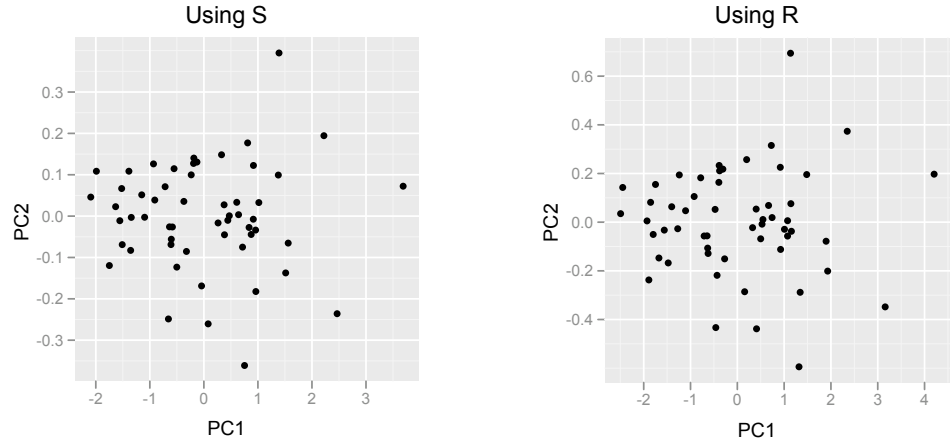
$$\mathbf{R} = \begin{bmatrix} 1.000 & 0.953 \\ 0.953 & 1.000 \end{bmatrix}$$

Variable	$\mathbf{e}_1$	$\mathbf{e}_2$
100m	0.707	0.707
200m	0.707	-0.707
Variance	1.95	0.0472
% Tot Var	97.6	100.0

- *Equal contribution from each variable. Why?*
- *PC 1 explains ___% of the variation. It drops, why?*
- *The correlation between PC 1 and the 100m is ___, and with the 200m is ____. Same as for S, why?*
- *Which analysis (on S or R) is more appropriate?*

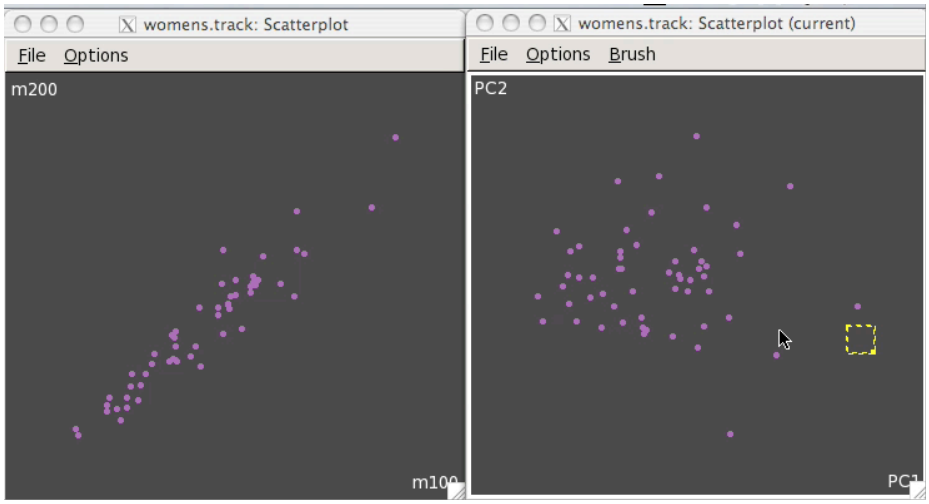
S vs R

Are they the same?



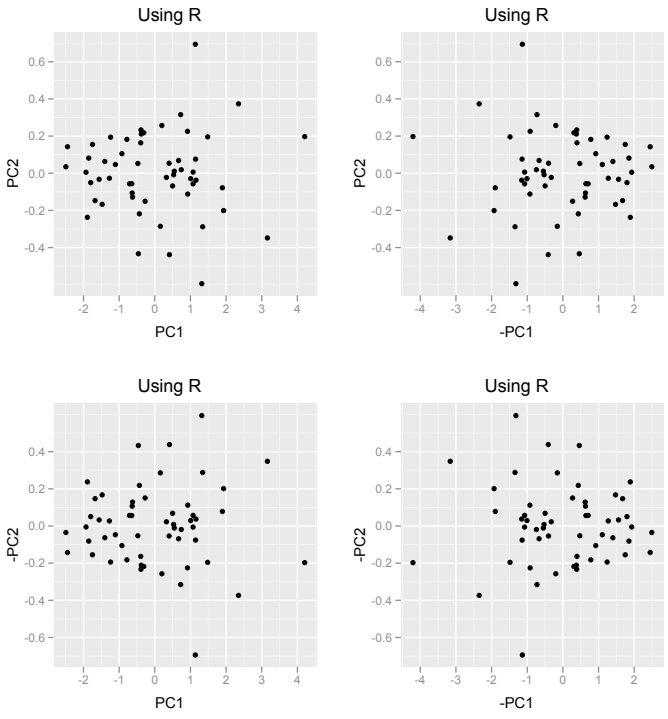
Exploring results

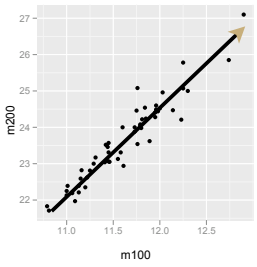
RAW *PCs*



Geometry

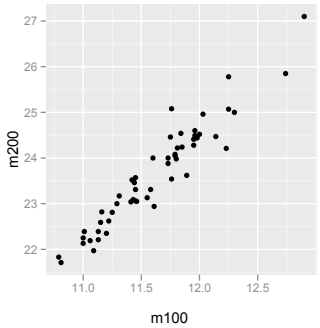
- *Are these the same?*
- _____
  _____
  _____.

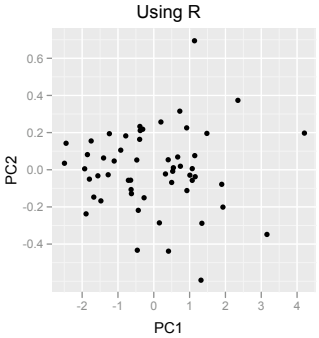




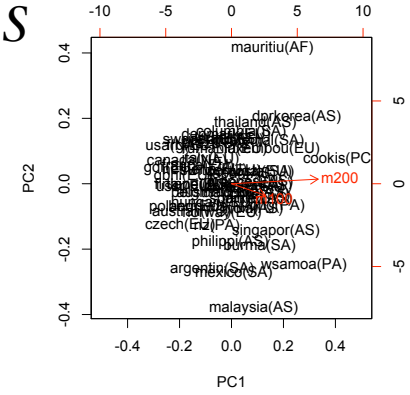
Geometry

Raw data is transformed by "sphering" to produce PCs.

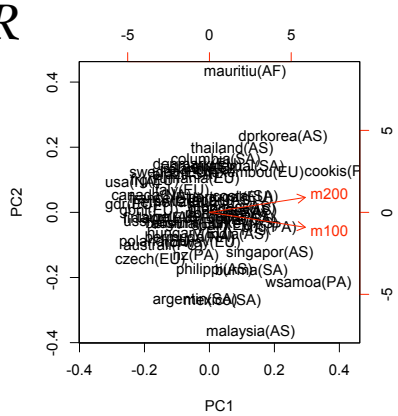




S



R



biplots

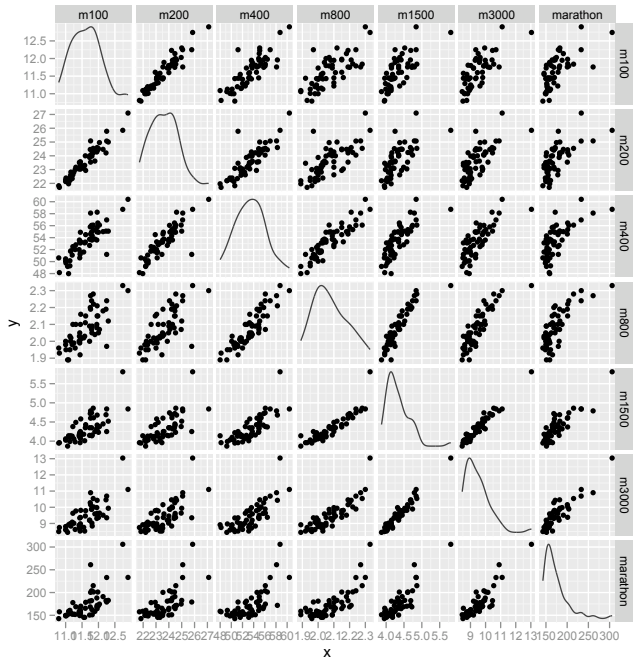
- *Original ____ plotted on the plot of the principal components.*
- *Helps to _____ the new variables.*
- *Here we learn that _____ _____ _____, slightly different weights for PCA on S, equal for PCA on R*

Example: Track records

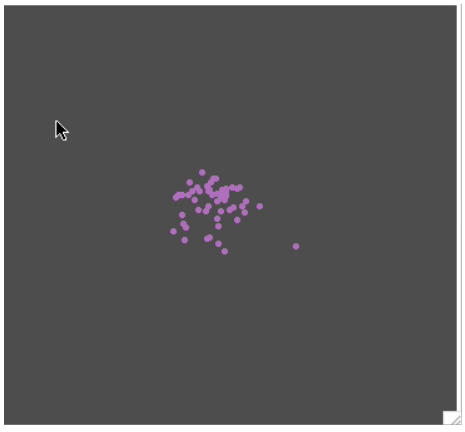
- *This data contains the women's national records for 100m, 200m, 400m, 800m, 1500m, 3000m and marathon, for 55 countries.*

country	m100 (sec)	m200 (sec)	m400 (sec)	m800 (min)	m1500 (min)	m3000 (min)	marathon (min)
argentin(SA)	11.61	22.94	54.50	2.15	4.43	9.79	178.52
australi(PC)	11.20	22.35	51.08	1.98	4.13	9.08	152.37
austria(EU)	11.43	23.09	50.62	1.99	4.22	9.34	159.37
...							

Pre-screening



- *What patterns do you see?*
- _____
  _____
  _____.



PCA on S or R?

S

	m100	m200	m400	m800	m1500	m3000	marathon
m100	0.204	0.479	1.01	0.036	0.109	0.276	9.4
m200	0.479	1.234	2.55	0.087	0.258	0.650	23.2
m400	1.011	2.550	7.17	0.260	0.701	1.717	57.5
m800	0.036	0.087	0.26	0.012	0.032	0.077	2.6
m1500	0.109	0.258	0.70	0.032	0.111	0.266	8.9
m3000	0.276	0.650	1.72	0.077	0.266	0.680	22.6
marathon	9.444	23.179	57.49	2.566	8.881	22.572	926.0

R

	m100	m200	m400	m800	m1500	m3000	marathon
m100	1.00	0.95	0.83	0.73	0.73	0.74	0.69
m200	0.95	1.00	0.86	0.72	0.70	0.71	0.69
m400	0.83	0.86	1.00	0.90	0.79	0.78	0.71
m800	0.73	0.72	0.90	1.00	0.90	0.86	0.78
m1500	0.73	0.70	0.79	0.90	1.00	0.97	0.88
m3000	0.74	0.71	0.78	0.86	0.97	1.00	0.90
marathon	0.69	0.69	0.71	0.78	0.88	0.90	1.00

- *Variances very _____ between variables.*
- *Variables measure in different _____.*
- *Should do PCA on ___.*

PCA on S

	$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_3$	$\mathbf{e}_4$	$\mathbf{e}_5$	$\mathbf{e}_6$	$\mathbf{e}_7$
100	0.0102	0.120	0.326	0.150	0.925	-.00166	0.0168
200	0.025	0.315	0.880	0.0140	-.354	0.025	0.012
400	0.062	0.934	-.328	-.122	0.013	-.022	-.025
800	0.003	0.026	-.0371	0.049	-.015	0.262	0.963
1500	0.019	0.039	-.055	0.340	-.034	0.900	-.265
3000	0.024	0.082	-.088	0.919	-.130	-.349	0.041
mar	0.997	-.070	-.002	-.020	0.002	-.000	-.000
Var	930.9	4.05	0.319	0.115	0.014	0.006	0.001
Cum%	99.5	99.8	99.9	99.9	99.9	100.0	100.0

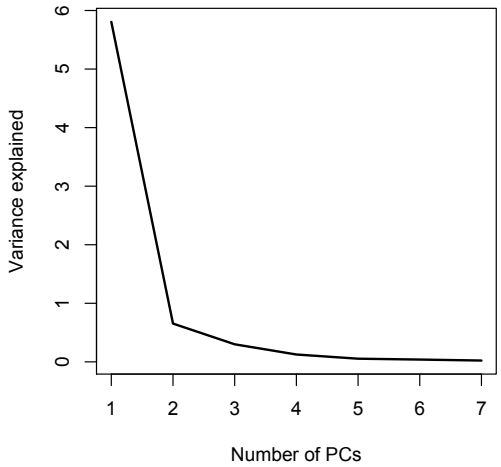
- *What's wrong with these results?*

PCA on R

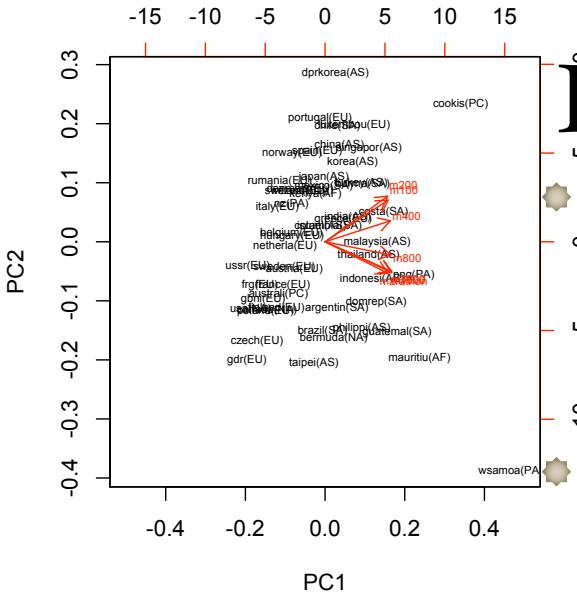
Variable	$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_3$	$\mathbf{e}_4$	$\mathbf{e}_5$	$\mathbf{e}_6$	$\mathbf{e}_7$
100	0.368	0.490	0.286	-.319	0.231	0.620	0.052
200	0.365	0.537	0.230	0.083	0.042	-.711	-.109
400	0.382	0.247	-.515	0.348	-.572	0.191	0.208
800	0.385	-.155	-.585	0.0421	0.620	-.019	-.315
1500	0.389	-.360	-.013	-.430	0.030	-.231	0.693
3000	0.389	-.348	0.153	-.363	-.463	0.009	-.598
mar	0.367	-.369	0.484	0.672	0.130	0.142	0.070
Var	5.81	0.654	0.300	0.125	0.054	0.039	0.022
Cum%	83.0	92.3	96.6	98.4	99.2	99.7	100.0

- *First PC is a linear combination of ___ events, _____ _____? It explains ____% of the variation.*
- *Second PC is a _____ between _____ and ____ distance events. Jointly explains ____%.*

How many PCs?



- *Scree plot: _____ vs _____*
- *Look for an "_____".*
- *Scree plot suggests _ PCs.*
- *Also consider _____ (_ and _), and proportion of total variance explained.*

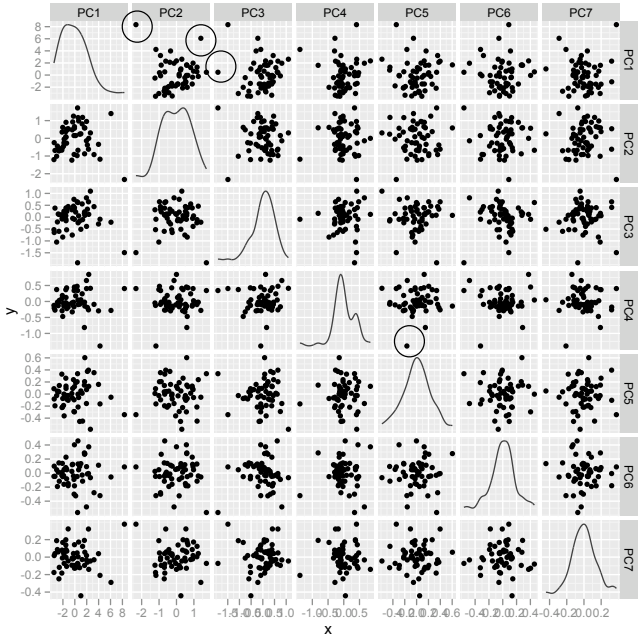


Interpretation

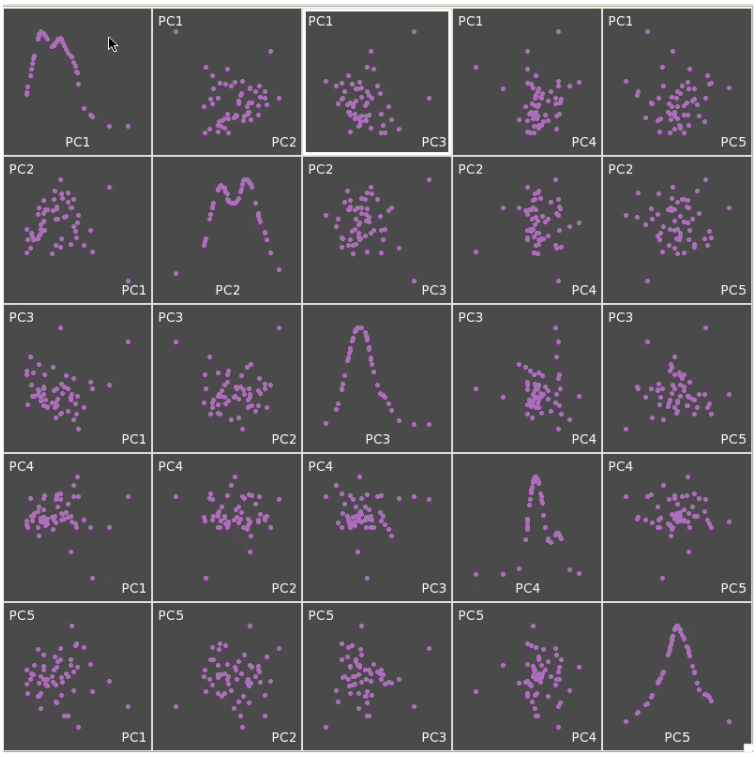
- *The first principal component is a roughly _____ combination of all of the variables. Consider PC1 to be a measure the the overall _____ of the athletic program.*
- *The second principal component is roughly a _____ between _____ distance and _____ distance events. Consider PC2 to be a measure of the countries skills in the _____ distance events.*

Which country has the best (worst) program? Short distance program?

Post-screening



- *Plot the PCs*
- *Should be ___ structure*
- *What do you see here?*
- _____
- *SOLUTION:* _____
  _____.



- *Outliers are:*

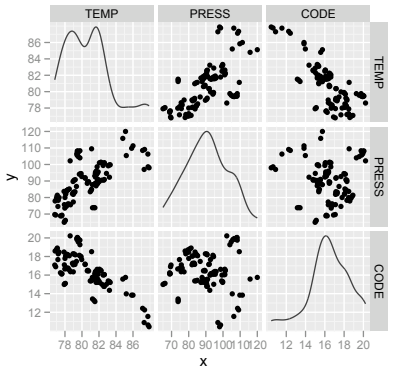
_____

_____

_____

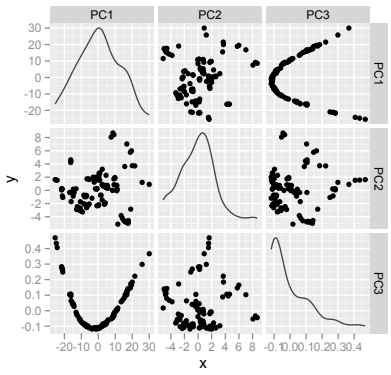
_____

More examples



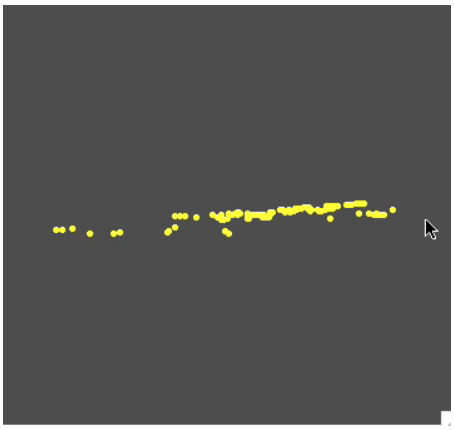
- *Flow through meters*

	Temp	Press	Code
Temp	1	0.63	-0.75
Press	0.63	1	0.04
Code	-0.75	0.04	1



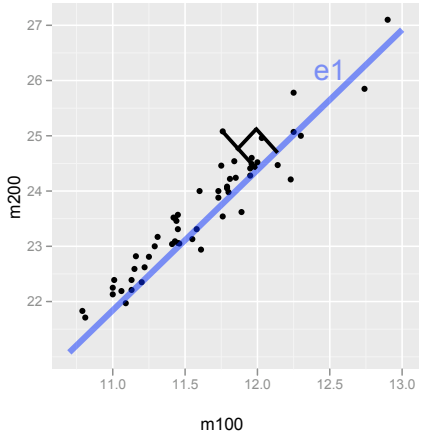
	PC1	PC2	PC3
Temp	0.14	0.7	-0.7
Press	0.99	-0.09	0.1
Code	0	-0.71	-0.7
Var	____	____	____
Prop	0.95	0.99	1

Flow meters



- *Classic example of _____: pairs of variables do not show extremely strong correlation.*
- *Relationship between the 3 variables is _____ defined by an equation though.*
- *PCA: Detects the main pattern of _____ association, removes this with PC1, PC2. The _____ association is captured in PC3.*

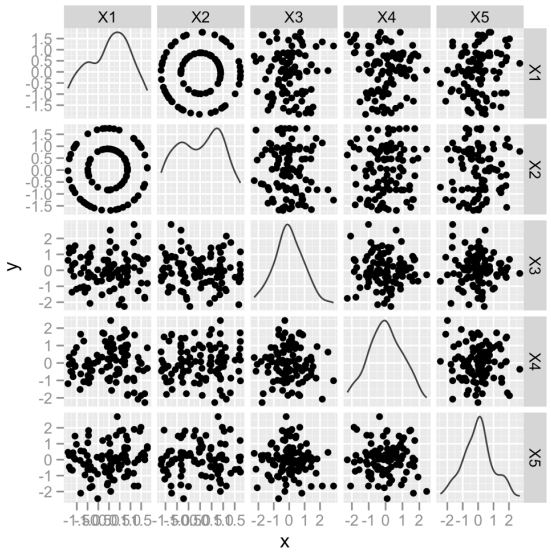
PCA vs Regression



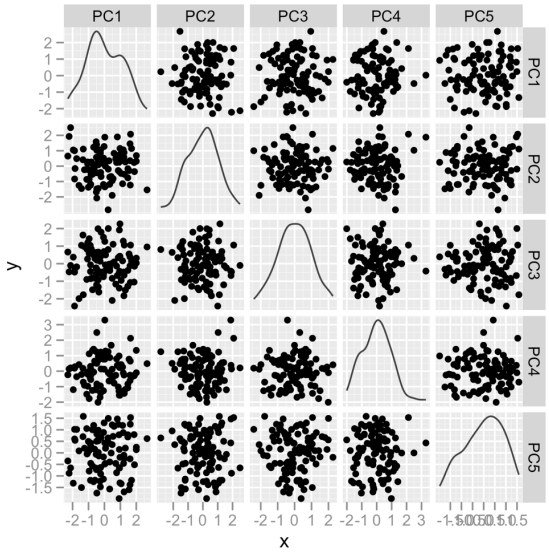
- *PCA is VERY MUCH LIKE _____.*
- *The first few PCs are like the _____ _____, and the lower PCs are like the _____. If there is non-linear dependence you'll likely see it in the lower PCs.*
- *The linear model is fit by minimizing _____ distance, not vertical distance between points and lines.*

More examples

Raw

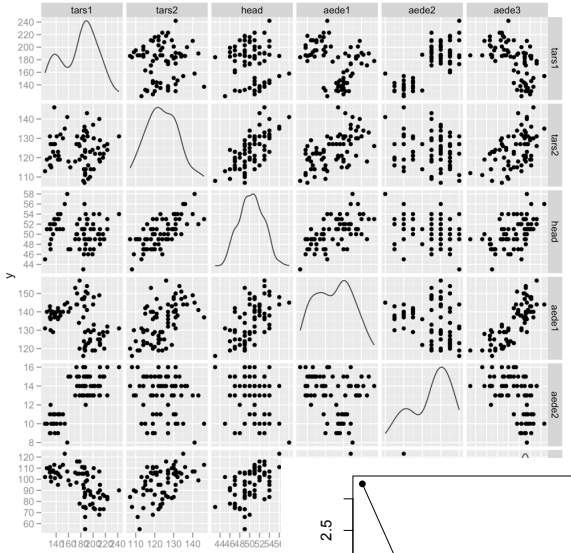


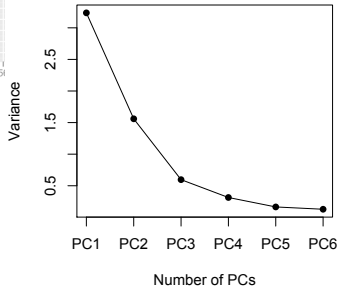
- *_____ structure is not recognized by PCA*



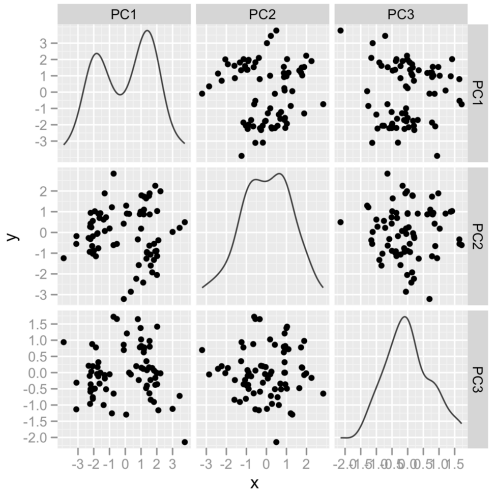
PCs

More examples





- *Flea beetles - _____ not revealed by PCA*



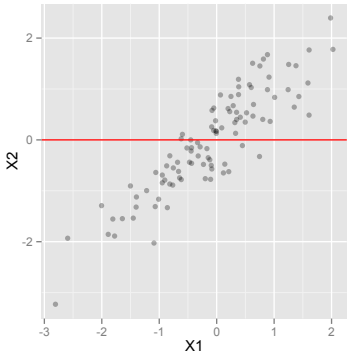
PP & PCA

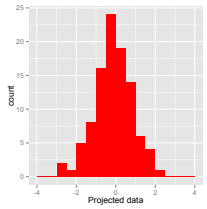
- *Projection pursuit _____ PCA. Any arbitrary function is optimized over all directions in the data.*
- *For PCA "f" is "_____"*

$$\max f(\mathbf{X}\mathbf{a}_1) \text{ subject to } \mathbf{a}_1'\mathbf{a}_1 = 1$$

Demo

- *2D data, look at variance of many possible projections*

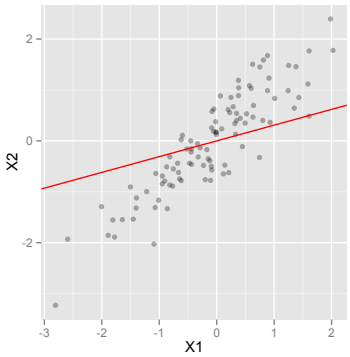


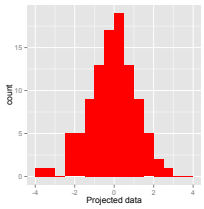


Var=0.92

Demo

- *2D data, look at variance of many possible projections*

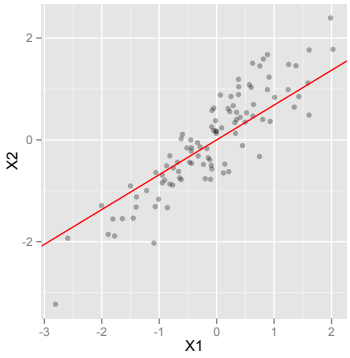


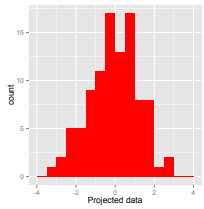


Var=1.4

Demo

- *2D data, look at variance of many possible projections*

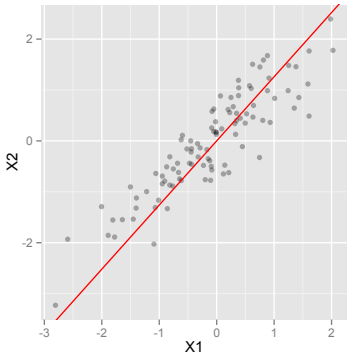


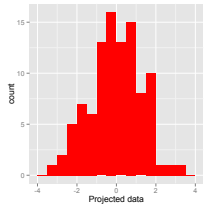


Var=1.7

Demo

- *2D data, look at variance of many possible projections*

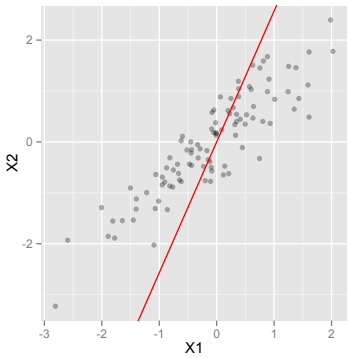


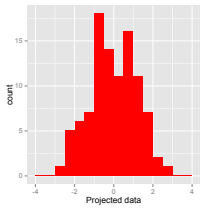


Var=1.8

Demo

- *2D data, look at variance of many possible projections*

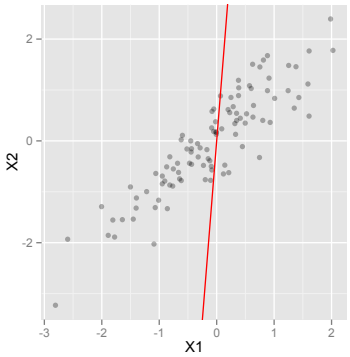


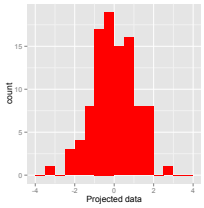


Var=1.6

Demo

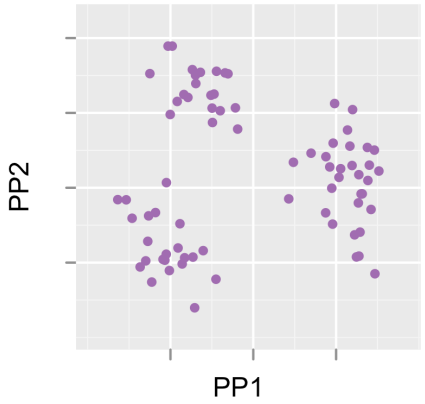
- *2D data, look at variance of many possible projections*





Var=1.1

PP & PCA



- *PP using the Holes index will find _____ that PCA misses.*

$$\mathbf{y} = \mathbf{X}\mathbf{A},$$

$$I_{\text{holes}}(\mathbf{A}) = \frac{1 - \frac{1}{n}\sum_{i=1}^{n}\exp(-\frac{1}{2}\mathbf{y}_i\mathbf{y}_i')}{1 - \exp(-\frac{p}{2})}$$

PCA & MDS

- *Multidimensional scaling (MDS) also _____ PCA.*
- *MDS finds a _____ of the data that _____ the interpoint distances, as closely as possible.*
- *PCA is MDS when _____ distance is used.*

Example: Womens track

Distances between all countries: variables standardized

	argentin(SA)	australi(PC)	austria(EU)	belgium(EU)	bermuda(NA)	brazil(SA)	burma(SA)
australi(PC)	2.738						
austria(EU)	2.357	0.983					
belgium(EU)	2.332	0.920	0.806				
bermuda(NA)	0.782	2.671	2.257	2.370			
brazil(SA)	1.112	2.100	1.673	1.846	0.736		
burma(SA)	1.916	4.041	3.372	3.305	2.257	2.571	
canada(NA)	3.231	0.737	1.510	1.412	3.095	2.547	4.525

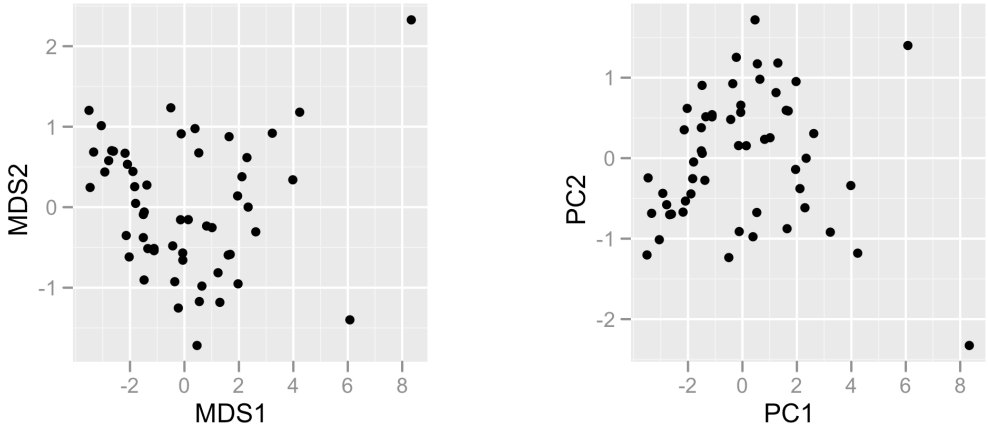
$D_{n \times n}$

- *MDS _____ the difference between these interpoint distances, and the distance between points in the low-dimensional representation.*

$$\text{Stress}_D(\mathbf{x}_1, \ldots \mathbf{x}_n) = \sqrt{\sum_{i \neq j} (D_{ij} - d_k(\mathbf{x}_i, \mathbf{x}_j))^2}$$

$d_k(\mathbf{x}_i, \mathbf{x}_j)$ is interpoint distance in k dimensions.

Example: Womens track



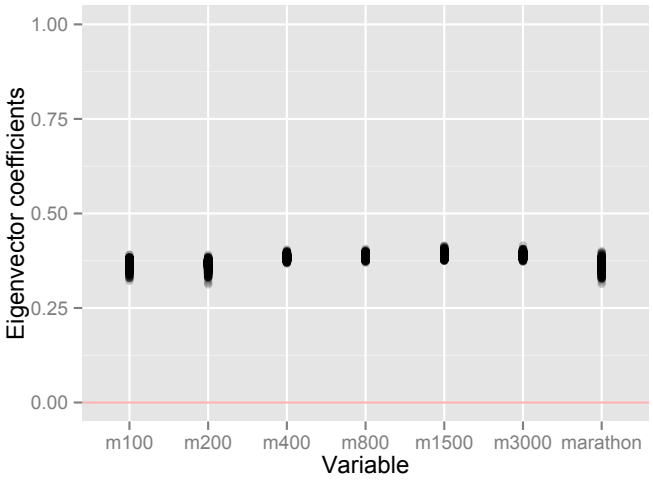
- *PCA is MDS when Euclidean distance is used.*
- *But _____ distance metrics can be used, and different _____ functions can produce _____ mappings.*

Practicalities

- *Womens track example, PC1=0.368 100m+0.365 200m + 0.382 400m+0.385 800m+0.389 1500m+ 0.389 3000m+ 0.367 marathon, where all variables have been standardized.*
- *What does this tell you???*
- _____

Bootstrap CI

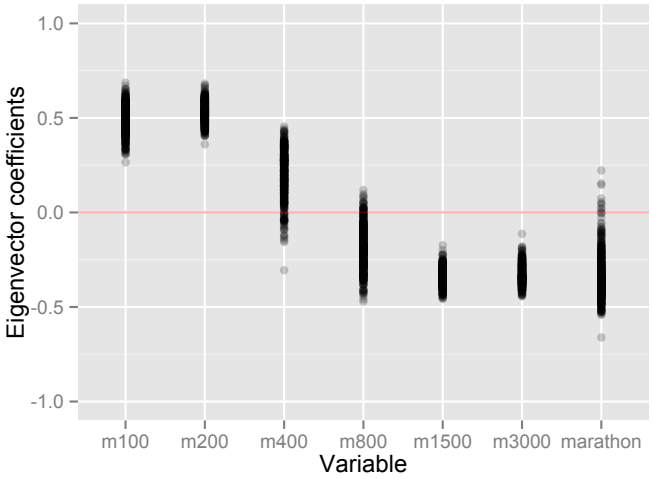
- *Bootstrap may be used to obtain some sense of the how close to zero an eigenvector coefficient is*
- *Bootstrap is _____ the data with replacement*
- *Compute the PCA, record the coefficients and repeat many times*



First PC

- *1000 samples*
- *All average around 0.37, which is an equal contribution from all variables*

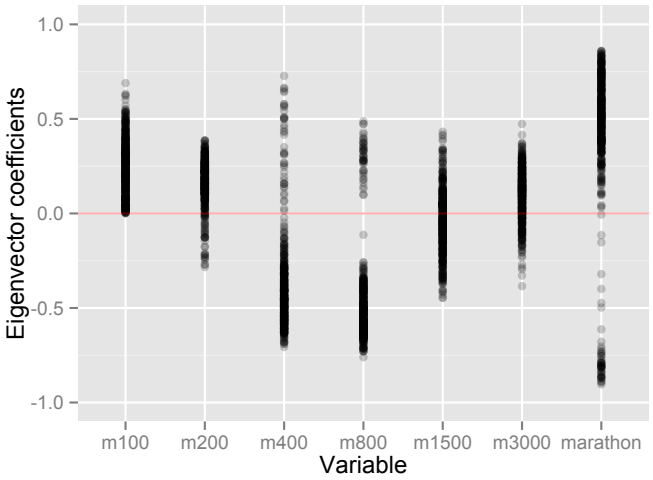
	m100	m200	m400	m800	m1500	m3000	marathon
lower	0.36	0.35	0.37	0.37	0.37	0.38	0.35
upper	0.39	0.39	0.39	0.39	0.40	0.40	0.39



2nd PC

- *1000 samples*
- *100m, 200m positive; 1500m, 3000m negative, others zero*

	m100	m200	m400	m800	m1500	m3000	marathon
lower	0.37	0.43	0.085	-0.373	-0.48	-0.48	-0.69
upper	0.66	0.65	0.561	0.091	-0.29	-0.28	-0.22



3rd PC

- *1000 samples*
- *All cross zero*

	m100	m200	m400	m800	m1500	m3000	marathon
lower	0.053	0.10	-1.61	-1.55	-0.31	-0.018	0.15
upper	0.558	0.66	-0.37	-0.46	0.34	0.582	1.80