Principal Component Analysis

- Uses:
  - Data Visualization
  - Data Reduction
  - Data Classification
  - Trend Analysis
  - Factor Analysis
  - Noise Reduction

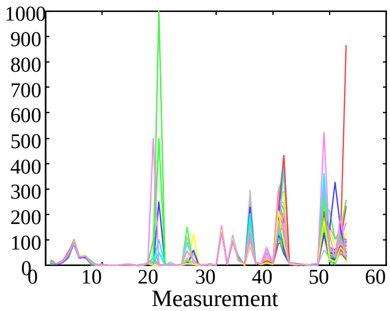
- Examples:
  - How many unique “sub-sets” are in the sample?
  - How are they similar / different?
  - What are the underlying factors that influence the samples?
  - Which time / temporal trends are (anti)correlated?
  - Which measurements are needed to differentiate?
  - How to best present what is “interesting”?
  - Which “sub-set” does this new sample rightfully belong?

Data Presentation

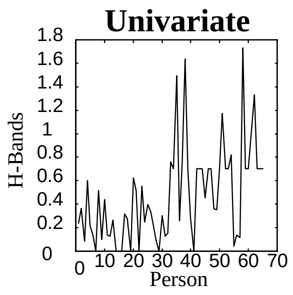
- Example: 53 Blood and urine measurements (wet chemistry) from 65 people (33 alcoholics, 32 non-alcoholics).
- Matrix Format

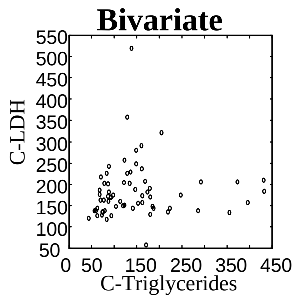
	H-WBC	H-RBC	H-Hgb	H-Hct	H-MCV	H-MCH	H-MCHC
A1	8.0000	4.8200	14.1000	41.0000	85.0000	29.0000	34.0000
A2	7.3000	5.0200	14.7000	43.0000	86.0000	29.0000	34.0000
A3	4.3000	4.4800	14.1000	41.0000	91.0000	32.0000	35.0000
A4	7.5000	4.4700	14.9000	45.0000	101.0000	33.0000	33.0000
A5	7.3000	5.5200	15.4000	46.0000	84.0000	28.0000	33.0000
A6	6.9000	4.8600	16.0000	47.0000	97.0000	33.0000	34.0000
A7	7.8000	4.6800	14.7000	43.0000	92.0000	31.0000	34.0000
A8	8.6000	4.8200	15.8000	42.0000	88.0000	33.0000	37.0000
A9	5.1000	4.7100	14.0000	43.0000	92.0000	30.0000	32.0000

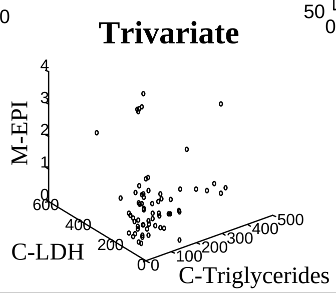
- Spectral Format



Data Presentation





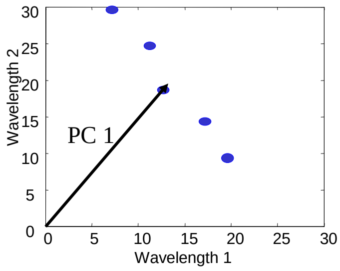


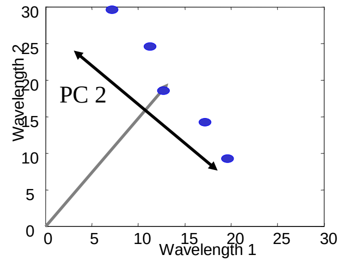
Data Presentation

- Better presentation than ordinant axes?
- Do we need a 53 dimension space to view data?
- How to find the ‘best’ low dimension space that conveys maximum useful information?
- One answer: Find Principal Components

Principal Components

- All principal components (PCs) start at the origin of the ordinant axes.
- First PC is direction of maximum variance from origin
- Subsequent PCs are orthogonal to 1st PC and describe maximum residual variance



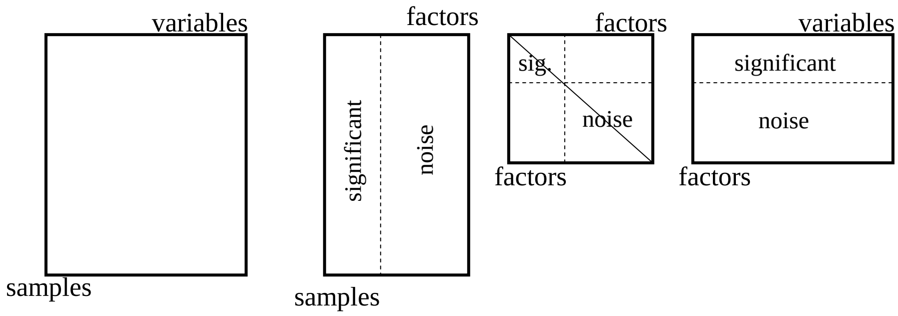


Principal Component Analysis

- Factor data, **R**, into 3 matrices.
  - $\mathbf{R}_{\text{samples x spectra}} = \mathbf{USV}^T$
- Columns of **V**
  - describe directions of maximum variance
  - linear combinations of ordinant spectral axes
  - are orthonormal
- Columns of **U**
  - describe relationship among samples
  - projection of each spectra onto column from **V**
  - are orthonormal
- Matrix **S**
  - Diagonal
  - Contains scale of **R**

Principal Component Analysis

$$\mathbf{R} = \mathbf{U} \quad \mathbf{S} \quad \mathbf{V}^T$$

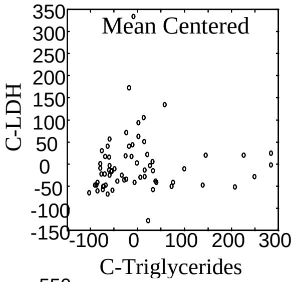


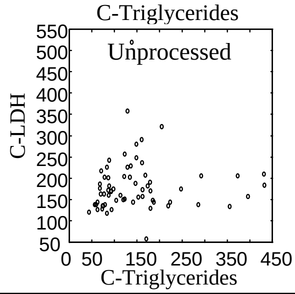
Principal Component Analysis

- Preprocessing
  - Data Translation
  - Data Scaling
- Axis Rotation
  - SVD
  - Varimax Rotation
- Determining Significant Factors
  - Statistical Tests
  - Empirical Tests
- Interpretation
  - Outlier Detection
  - Variable Selection

Preprocessing

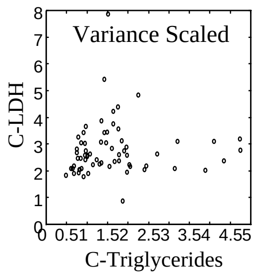
- Mean Centering
  - Translates center of data cloud to origin
  - For $\mathbf{R}_{(I \times J)}$, subtract mean response of the *I* samples from each of the *J* variables.
  - $R_{ij} = R_{ij} - r_j$
  - Rmc = R - (ones(I,1)* mean(R))

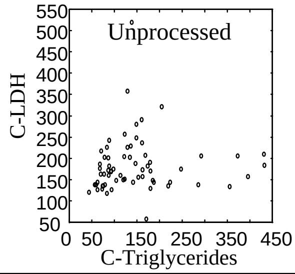




Preprocessing

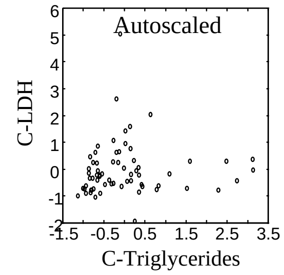
- Variance Scaling
  - Normalize each axis to same Euclidean length. Each variable will have same least-squares weight.
  - For $\mathbf{R}_{(I \times J)}$, subtract mean response of the *I* samples from each of the *J* variables.
  - $R_{ij} = R_{ij} / s_j$
  - Rvs = R - (ones(I,1) * std(R))

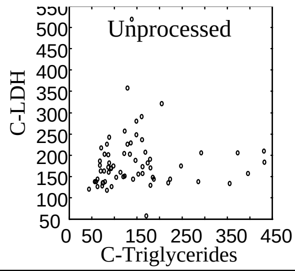




Preprocessing

- Autoscaling
  - Translates and stretches axes such that each variables has a mean 0 and standard deviation 1.
  - First mean center than variance scale the data





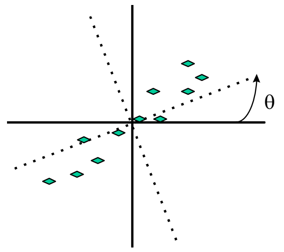
Preprocessing

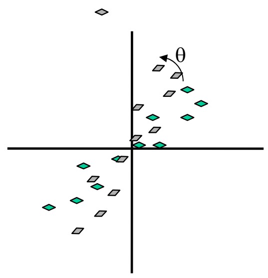
- Unit Area
  - Area under each spectrum is set to 1. Adjusts for changes in sampling size and lamp intensity
  - Divide each sample by sum of the measurements in each sample
  - $R_{ij} = R_{ij} / S_{j=1} R_{ij}$
  - Rua = R./(sum(R')'*ones(1,J))

- Unit Length
  - Length of each spectrum is set to 1. Places all data on unit circle
  - Divide each sample by sum of squared values of the measurements in each sample
  - $R_{ij} = R_{ij} / S_{j=1} R_{ij}^2$
  - Rul = R./(sum(R'.^2)'*ones(1,J))

Rotation of Axes

- Consider
  - $\mathbf{R}_{new} = \mathbf{R}Q$
  - $\mathbf{R}_{new} = \mathbf{R}Q^T$
  - $$\Theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$
- Q is a rotation matrix
  - $\mathbf{R}_{new} = \mathbf{R}Q$ Rotates Axes
  - $\mathbf{R}_{new} = \mathbf{R}Q^T$ Rotates Data
- $\mathbf{R}_{new}$ is location of data on new axes



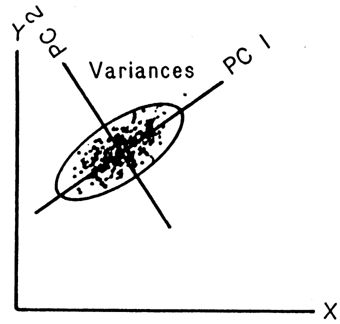


Axis Rotation

- Goal: Find new axis set, **Q**, such that:
  - 1. All axes are orthogonal (orthonormal)
  - 2. The sum of the squared distances from all points to the first axis is minimized: $||\mathbf{R} - \mathbf{Q}_1\mathbf{Q}_1^T\mathbf{R}||$
  - 3. Given $\mathbf{Q}_{n-1}$ exist, find $\mathbf{Q}_n$ such that 1 and 2 are satisfied and $||\mathbf{R} - \mathbf{Q}_{1:n}\mathbf{Q}_{1:n}^T\mathbf{R}||$ is minimized.
- In other words….
  - Now, $\mathbf{X} = \mathbf{X}\mathbf{I}^T$ where columns of **I** are orthonormal axes (directions) and rows in **X** are distances along each unit axes.
  - Want, $\mathbf{X} = \mathbf{P}\mathbf{Q}^T$ where columns of **Q** are orthonormal axes (directions) satisfying 1 - 3 and rows in **P** are distances along the new axes.
  - Eventually want: $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ where $\mathbf{V}=\mathbf{Q}$ and $\mathbf{U}\mathbf{S} = \mathbf{P}$ with columns of **V** and **U** being orthonormal and **S** being a diagonal matrix.

Three methods for Rotation

- Eigenvalue Problem (EP)
- Nonlinear Iterative Partial Least Squares (NIPLS)
- Singular Value Decomposition (SVD)



Eigenvector Rotation

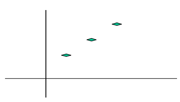
- Given,
  - **C** = $\mathbf{R}^T\mathbf{R}$ which is a correlation matrix if **R** is mean centered and unit length.
  - $\mathbf{R}_{new} = \mathbf{RQ}$ and
  - $\mathbf{R}_{new}^T = \mathbf{Q}^T\mathbf{R}^T$
- Want
  - $\mathbf{R}_{new}^T\mathbf{R}_{new} = \mathbf{Q}^T\mathbf{R}^T\mathbf{RQ} = \mathrm{Q}^T\mathbf{CQ} = \mathbf{L}$
  - where **L** is diagonal
- This implies that
  - $\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}^{-1}\mathbf{Q}$
  - so, **CQ** = **QL**
- This sets up an eigenvalue problem
  - **CQ** = **QL**

Eigenvector Rotation

- The EP, **CQ** = **QL**, can be solved 1 l at a time:
  - **CQ** = l**Q**
  - **CQ** - l**Q** = 0
  - (**C** - l**I**)**R** = 0
- Solution only exists if **|C** - l**I|** = 0
- So, solve for roots of l then solve for columns of **R**

Eigenvector Rotation Example

$$\mathbf{R} = \begin{bmatrix} 4 & 2 \\ 6 & 4 \\ 8 & 6 \end{bmatrix}$$



Autocale **R**:

Mean Center:

$$\bar{\mathbf{r}}_1 = 3; \quad \bar{\mathbf{r}}_2 = 2; \quad \mathbf{R} = \begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}$$

Variance Scale:

$$\|\mathbf{r}_1\| = \sqrt{2}; \quad \|\mathbf{r}_2\| = \sqrt{2}; \quad \mathbf{R} = \begin{bmatrix} -1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

Find $\mathbf{C} = \mathbf{R}_{vs}{}^{T}\mathbf{R}_{vs}$

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

Diagonalize **C**

$$|\mathbf{C} - \lambda\mathbf{I}| = \begin{vmatrix} 1-\lambda & 1 \\ 1 & 1-\lambda \end{vmatrix} = 0$$

Eigenvector Rotation Example

- Solve for roots
  - $(1 - l)^2 - 1 = 0$
  - $l\,(l-2) = 0$
  - $l_1 = 2,\ l_2 = 0$
  - Note $\sum \lambda_k = Trace[\mathbf{C}]$
- Solve for first eigenvector (l = 2):

$$\begin{bmatrix} 1-2 & 1 \\ 1 & 1-2 \end{bmatrix}\begin{bmatrix} v_{11} \\ v_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

  - $-v_{11} + v_{21} = 0$
  - $v_{11} - v_{21} = 0$
  - \ $v_{11} = v_{21}$
- Normalize 1st eigenvector to unit length

$$\mathbf{v}_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

- Solve for 2nd eigenvector (l = 0):

$$\begin{bmatrix} 1-0 & 1 \\ 1 & 1-0 \end{bmatrix}\begin{bmatrix} v_{12} \\ v_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

  - $v_{12} + v_{22} = 0$
  - $v_{12} + v_{22} = 0$
  - \ $v_{12} = -v_{22}$

Eigenvector Rotation Example

- Normalize 2nd eigenvector to unit length:

$$\mathbf{v}_2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} \text{ or } \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

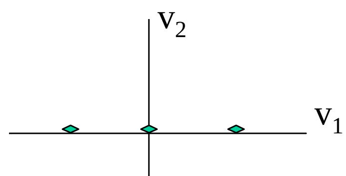
- Therefore

$$\mathbf{V} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

- Find points on new axis

$$\mathbf{R}new = \begin{bmatrix} -1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}$$

- Graphically



- Note: $\mathbf{v}_1$ contains 100% of the information

PCA EP Summary

- Translation
  - center and scale **R**
- Covariance matrix
  - $\mathbf{C} = \mathbf{R}^T\mathbf{R}$
- Diagonalize **C**

$$\left|\mathbf{C} - \lambda\,\mathbf{I}\right| = \begin{vmatrix} 1-\lambda & 1 \\ 1 & 1-\lambda \end{vmatrix} = 0$$

- Find rotation matrix
  - $(\mathbf{C}-l\mathbf{I})\mathbf{v} = 0$
  - $\mathbf{V}=[\mathbf{v}_1| \ldots |\mathbf{v}_n]$
- Score matrix:
  - $\mathbf{R}_{new} = \mathbf{RV}$

PCA by NIPLS

- NIPLS is an iterative methods where 1 factor is calculated at a time.
- Uses model
  - $\mathbf{R} = \mathbf{PQ}^T$
- 1. Start with normalized guess of scores for n^{th} factor **p** where $||\mathbf{p}^T\mathbf{p}|| = 1$
- 2. Calculate R-block loadings, **q**

$$\mathbf{q} = \frac{\mathbf{R}^T\mathbf{p}}{\left\|\mathbf{p}^T\mathbf{R}\mathbf{R}^T\mathbf{p}\right\|}$$

- 3. Calculate new X-block score vector, **p**
  - $\mathbf{p} = \mathbf{Rq}$
- 4. Check for convergence,
  - abs($||\mathbf{p}^T\mathbf{p}||_{new}$ - $||\mathbf{p}^T\mathbf{p}||_{old}$) < 10^{-6},
  - If not converged loop back to 2.
- 5. If converged, variance (information) described by this PC is subtracted from **R**
  - $\mathbf{R}_{new} = \mathbf{R} - \mathbf{pq}^T$
- 6. If more PCs are needed, return to 1 with $\mathbf{R}_{new}$ (from 5).

PCA by Singular Value Decomposition

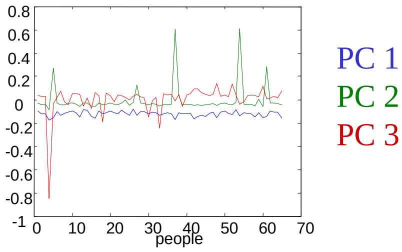
- $\mathbf{X}_{(IxJ)}$ has a singular value decomposition
- $\mathbf{X}_{(IxJ)} = \mathbf{U}_{(IxI)}\ \mathbf{S}_{(IxJ)}\ \mathbf{V}_{(JxJ)}^{T}$
  - where
    - $\mathbf{U}^T\mathbf{U} = \mathbf{I}_{I,I}$
    - $\mathbf{V}^T\mathbf{V} = \mathbf{I}_{J,J}$
  - $\mathbf{U}$ are eigenvectors of $\mathbf{XX}^T$
  - $\mathbf{V}$ are eigenvectors of $\mathbf{X}^T\mathbf{X}$
  - $\mathbf{S}$ are square roots of eigenvalues from $\mathbf{XX}^T$ or $\mathbf{X}^T\mathbf{X}$
- Proof:
  - $\mathbf{X}=\mathbf{USV}^T$ and $\mathbf{X}^T=\mathbf{VSU}^T$
  - $\mathbf{X}^T\mathbf{X} = \mathbf{VSU}^T\mathbf{USV}^T$
  - $\mathbf{X}^T\mathbf{X} = \mathbf{VS}^2\mathbf{V}^T$
  - $\mathbf{X}^T\mathbf{XV} = \mathbf{VS}^2$
- Pseudoinverse:
  - $\mathbf{X}^+ = (\mathbf{USV}^T)^+ = \mathbf{VS}^{-1}\mathbf{U}^T$
- Projection:
  - $\mathbf{X}^+\mathbf{X} = \mathbf{VS}^{-1}\mathbf{U}^T\mathbf{USV}^T = \mathbf{VV}^T$
  - $\mathbf{XX}^+ = \mathbf{USV}^T\ \mathbf{VS}^{-1}\mathbf{U}^T = \mathbf{UU}^T$

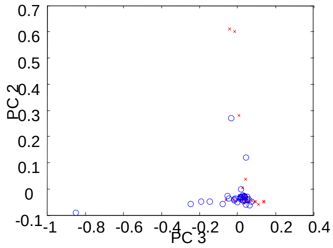
Notes Regarding Factor Analysis

- Any matrix can be factored into
  - $\mathbf{R} = \mathbf{PQ}^T$
  - $\mathbf{R} = \mathbf{USV}^T$
- Infinite ways to factor **R**
- PCA (orthogonal factors) is just one possibility
- Not all factors are meaningful
  - just mathematical constructs
- But, some factors (models) are useful!
- Determining ‘significant’ factors is difficult.
- Noise filtering is possible by eliminating ‘non-significant’ factors
- Can be used for data and variable reduction

Example

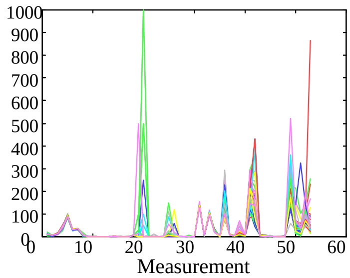
- Variables: 53 blood and urine measurements
- Objects: 65 people (33 “alcoholics” in treatment; 32 “non-alcoholics”)
- PCA on un-normalized data
  - Why doesn’t it work?





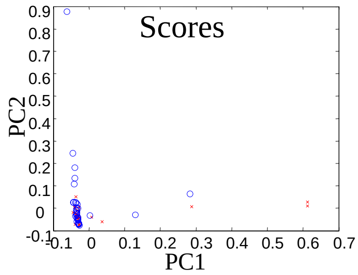
Example

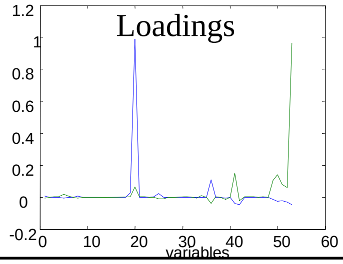
- Look at raw data
  - some variables have large mean
  - others have large variance
  - this variance does not translate into ‘predictive’ variance
  - some variables have very low variance
  - these may be more useful
  - but are swamped by other variances



Example

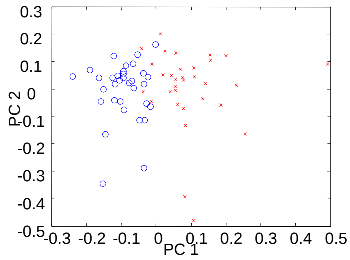
- PCA on mean centered data
- Still no classification
- Why?
  - Look at loadings
  - A couple of high variance measurements dominate the data set

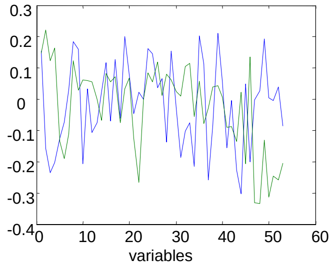




Example

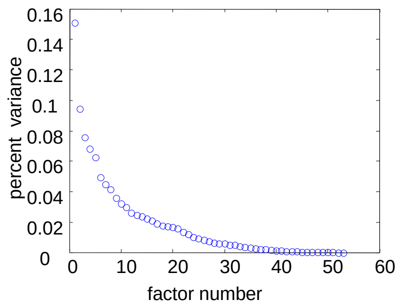
- PCA on autoscaled data
- Decent classification
  - only PC1 is useful
- Loadings not dominated by any one variable
- Questions
  - How much variance does each PC describe?
  - Are all variables needed?
  - Any outliers?





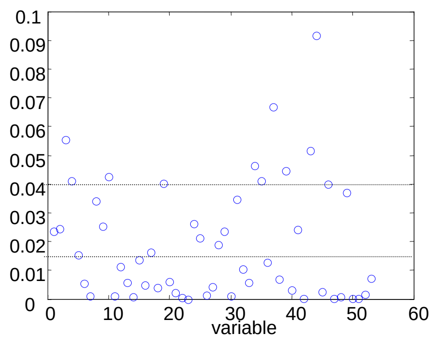
Example

- Factored data **R** = **USV**^{T}
- Recall that scale is in **S**
- Total variance SS_{ii}^2
- Percent variance captured by *i*th PC is
  - S_{ii}^2 / SS_{ii}^2



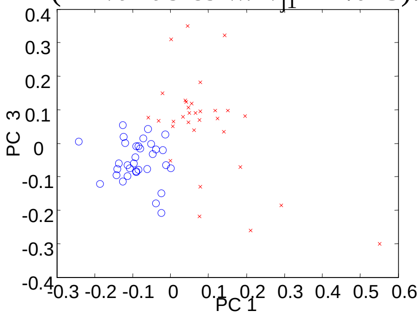
Example

- Squaring a loading gives an estimate of the leverage each variable has on determining the loading.
- Variables with small leverage can be eliminated with little impact on the overall model

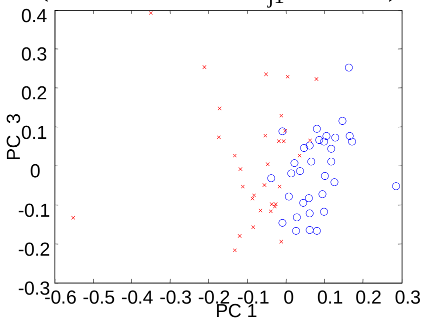


Example

PCA on autoscaled data
(24 variables w/ $v_{j1}^2 > .015$):



PCA on autoscaled data
(10 variables w/ $v_{j1}^2 > .040$):



Determining Number of PC

- No single best method
- Subjective
- Knowledge of data and errors are helpful

- Some tactics:
  - Variance described
  - Inspection of scores and loadings
  - Experimental error
  - Empirical
  - Statistical
  - Cross validation