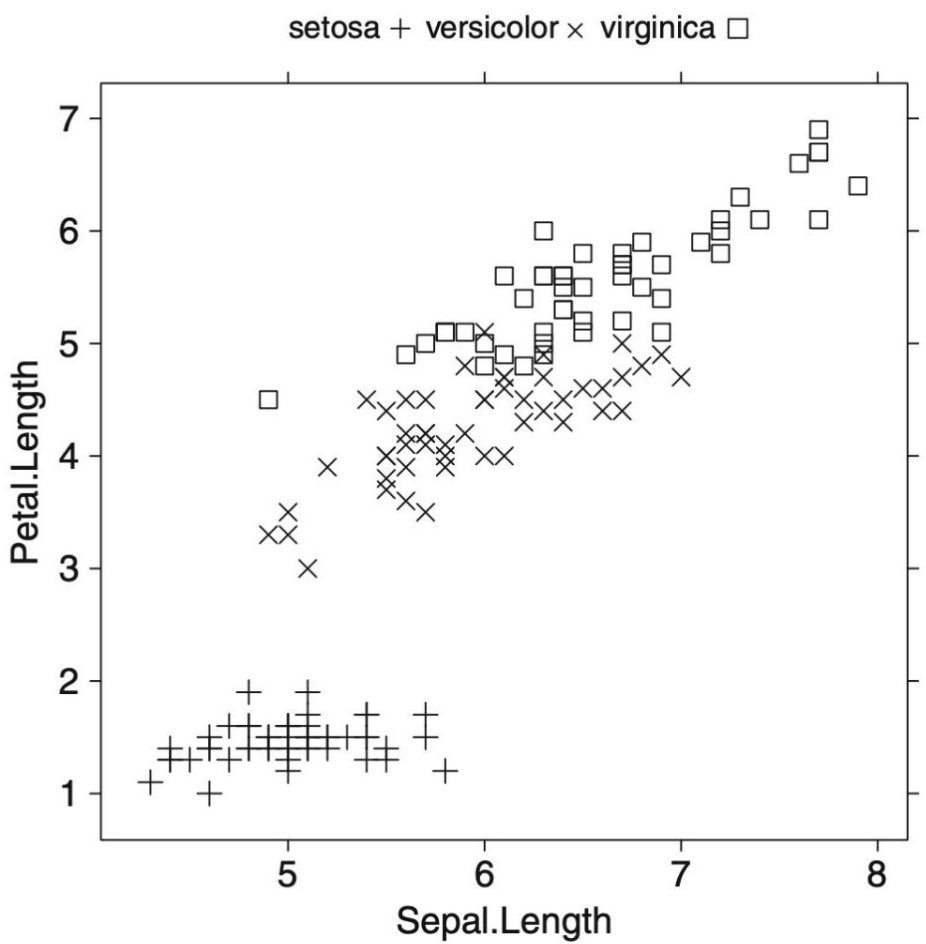
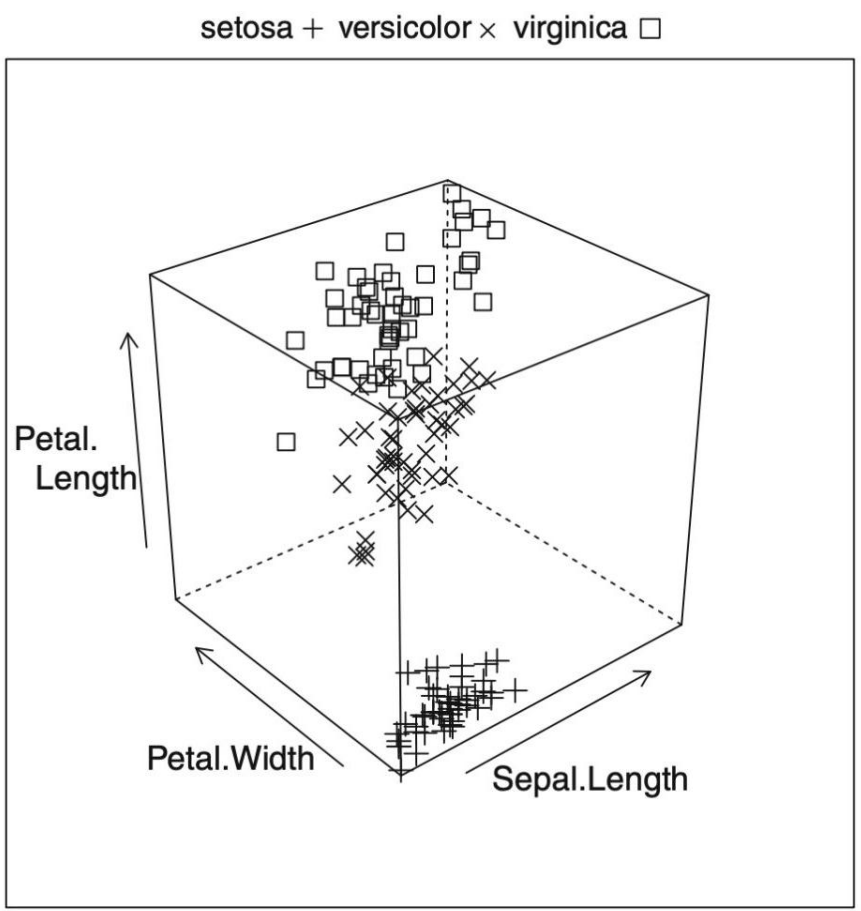
Computing for Medicine

Data Science: PCA and Feature Selection

Tavpritesh Sethi

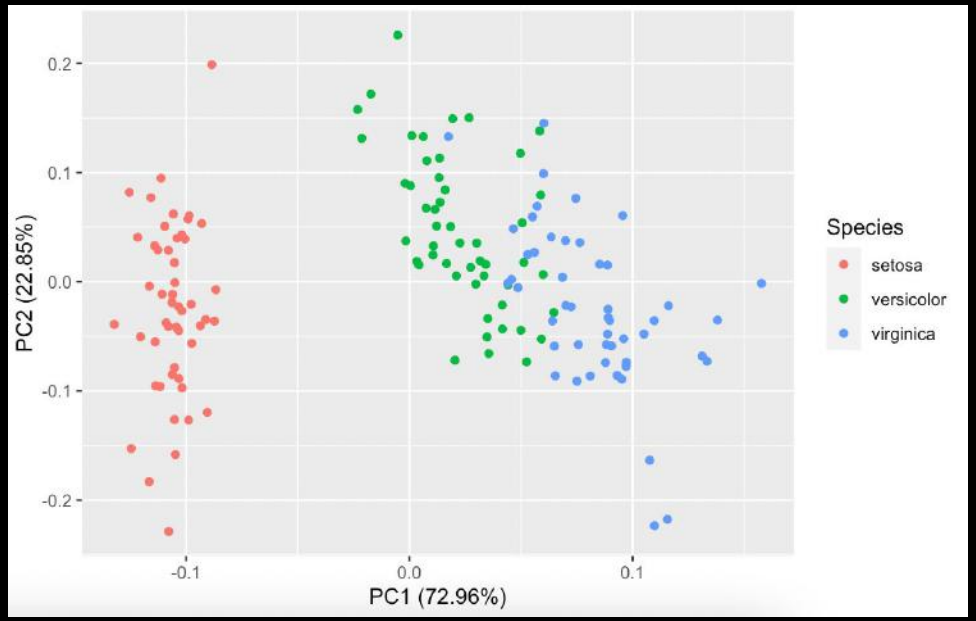
High Dimensional Data
Principal Component Analysis


setosa + versicolor × virginica □
Petal.Length
Sepal.Length
1
2
3
4
5
6
7
5
6
7
8

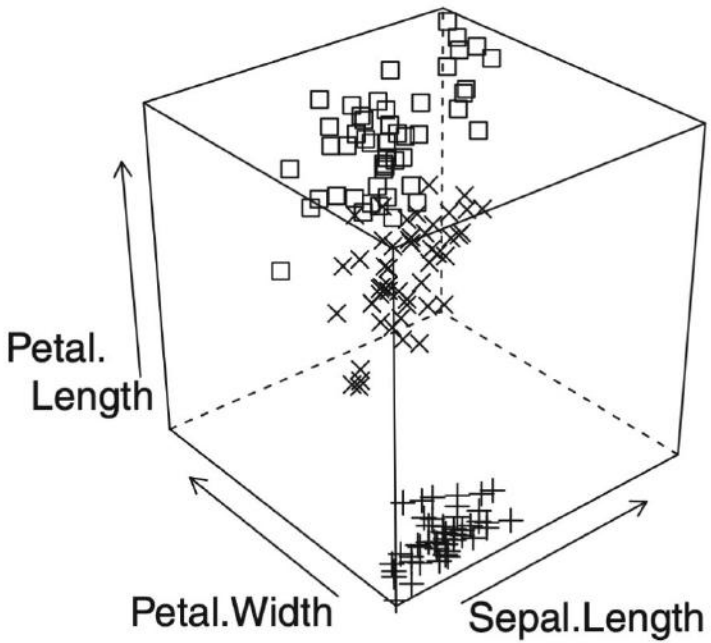

setosa + versicolor × virginica □
Petal.
Length
Petal.Width
Sepal.Length

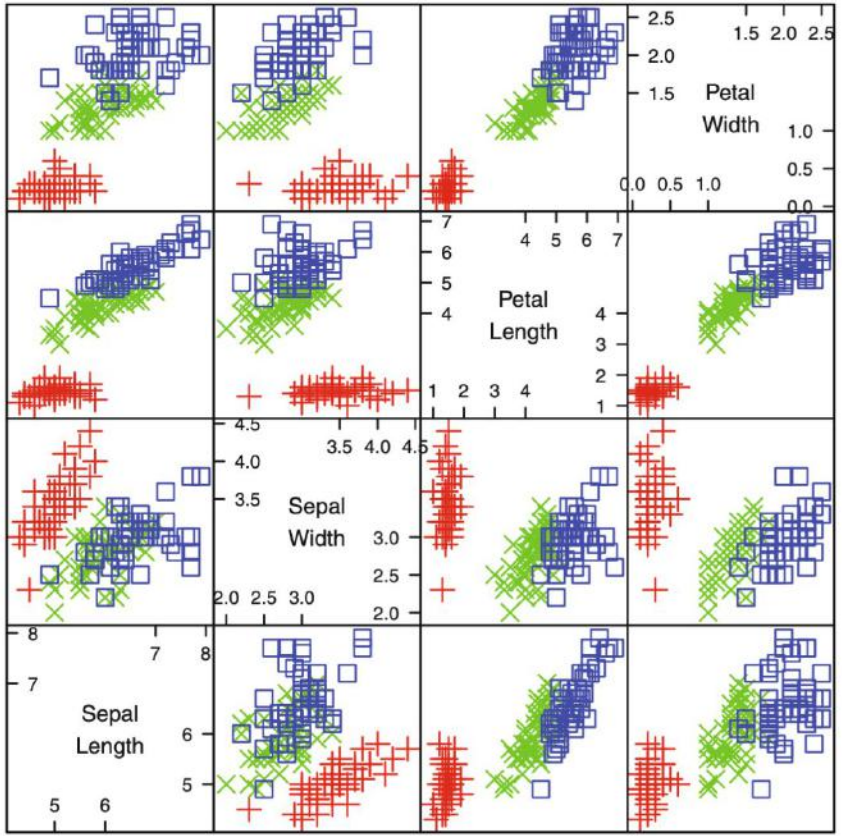
```r
library(ggfortify)
df <- iris[1:4]
pca_res <- prcomp(df, scale. = TRUE)

autoplot(pca_res)
autoplot(pca_res, data = iris, colour = 'Species')
```



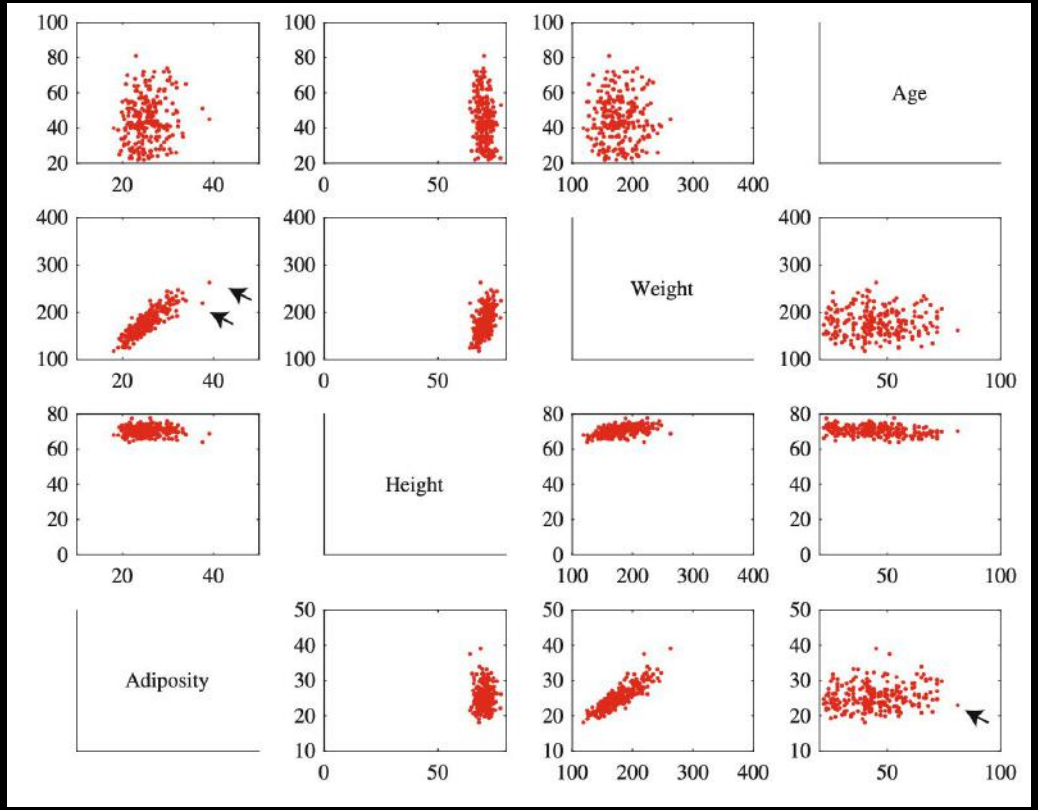
setosa + versicolor × virginica □





Scatter Plot Matrix

Height-Weight Dataset: Original Scatterplots



Challenges of High Dimensional Data

Remember This: *High dimensional data does not behave in a way that is consistent with most people's intuition. Points are always close to the boundary and further apart than you think. This property makes a nuisance of itself in a variety of ways. The most important is that only the simplest models work well in high dimensions.*

Recap: Covariance and Correlation

$$\mathsf{cov}\left(\{x\},\{y\}\right) = \frac{\sum_i (x_i - \mathsf{mean}\left(\{x\}\right))(y_i - \mathsf{mean}\left(\{y\}\right))}{N}$$

$$\mathsf{corr}\left(\{(x,y)\}\right) = \frac{\mathsf{cov}\left(\{x\},\{y\}\right)}{\sqrt{\mathsf{cov}\left(\{x\},\{x\}\right)}\sqrt{\mathsf{cov}\left(\{y\},\{y\}\right)}}.$$

Covariance Matrix

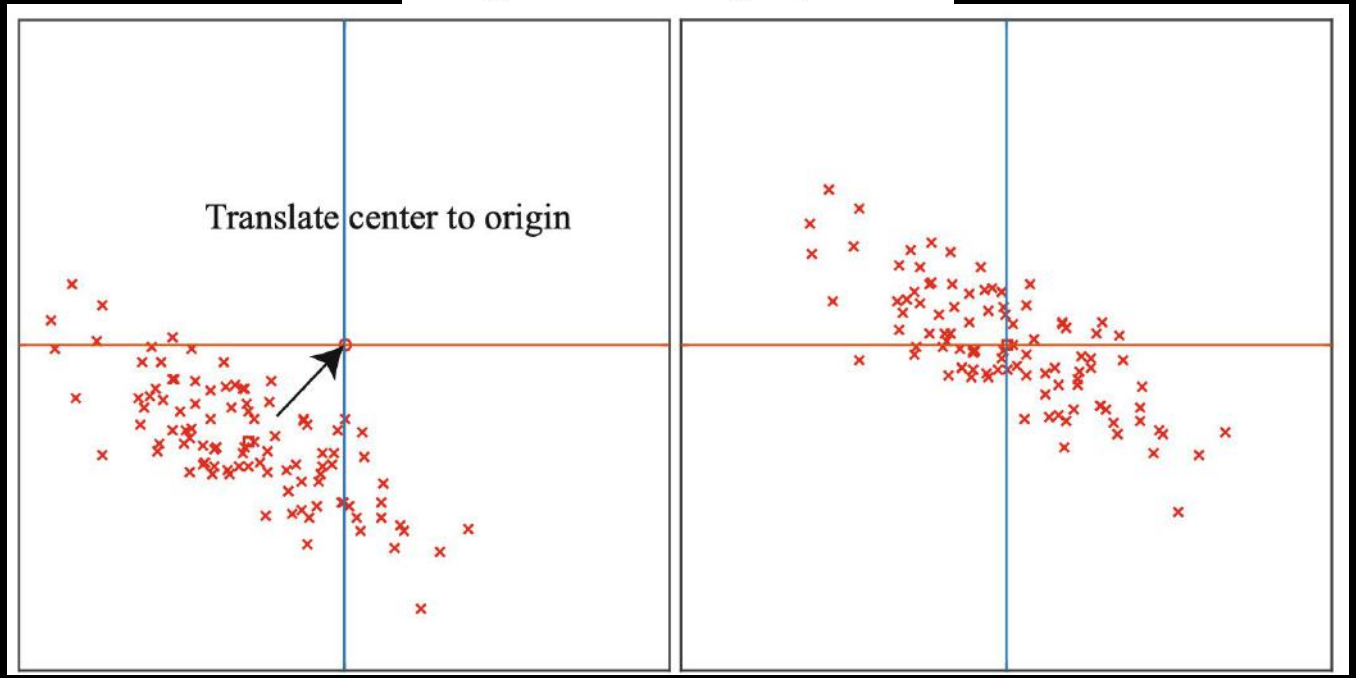
$$\text{Covmat}\left(\{\mathbf{x}\}\right) = \frac{\sum_i (\mathbf{x}_i - \text{mean}\left(\{\mathbf{x}\}\right))(\mathbf{x}_i - \text{mean}\left(\{\mathbf{x}\}\right))^T}{N}$$

Useful Facts: 4.3 *Properties of the Covariance Matrix*

- The j, k'th entry of the covariance matrix is the covariance of the j'th and the k'th components of $\mathbf{x}$, which we write $\text{cov}\left(\left\{x^{(j)}\right\}, \left\{x^{(k)}\right\}\right)$.
- The j, j'th entry of the covariance matrix is the variance of the j'th component of $\mathbf{x}$.
- The covariance matrix is symmetric.
- The covariance matrix is always positive semidefinite; it is positive definite, *unless* there is some vector $\mathbf{a}$ such that $\mathbf{a}^T(\mathbf{x}_i - \text{mean}\left(\{\mathbf{x}_i\}\right) = 0$ for all i.

Affine Transformation

$$\mathbf{m}_i = \mathcal{A}\mathbf{x}_i + \mathbf{b}.$$



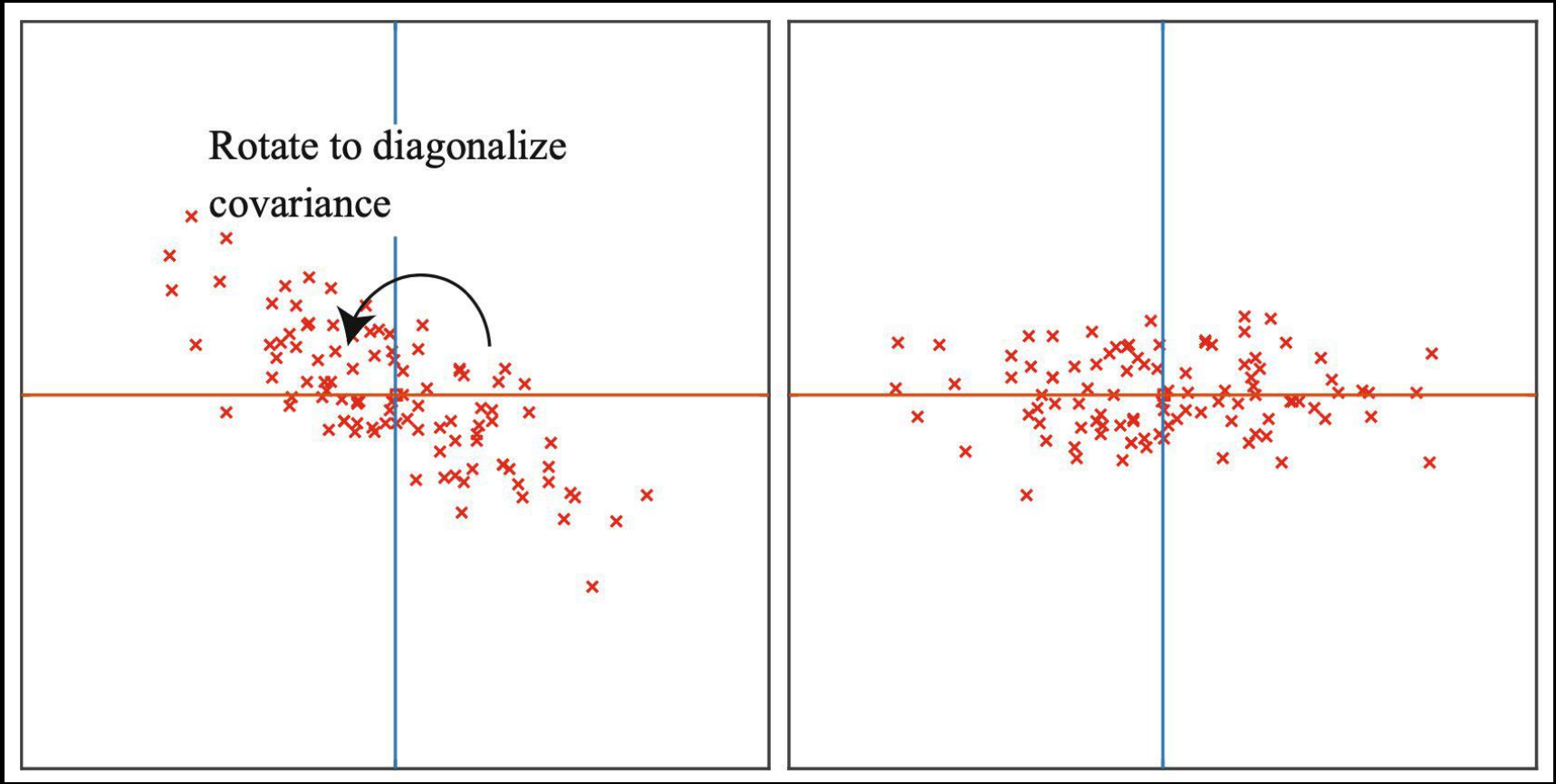
Mean and Covariance matrix under Affine Transformations

$$\begin{aligned}\text{mean}\left(\{\mathbf{m}\}\right) &= \text{mean}\left(\{\mathcal{A}\mathbf{x}+\mathbf{b}\}\right)\\ &= \mathcal{A}\text{mean}\left(\{\mathbf{x}\}\right)+\mathbf{b},\end{aligned}$$

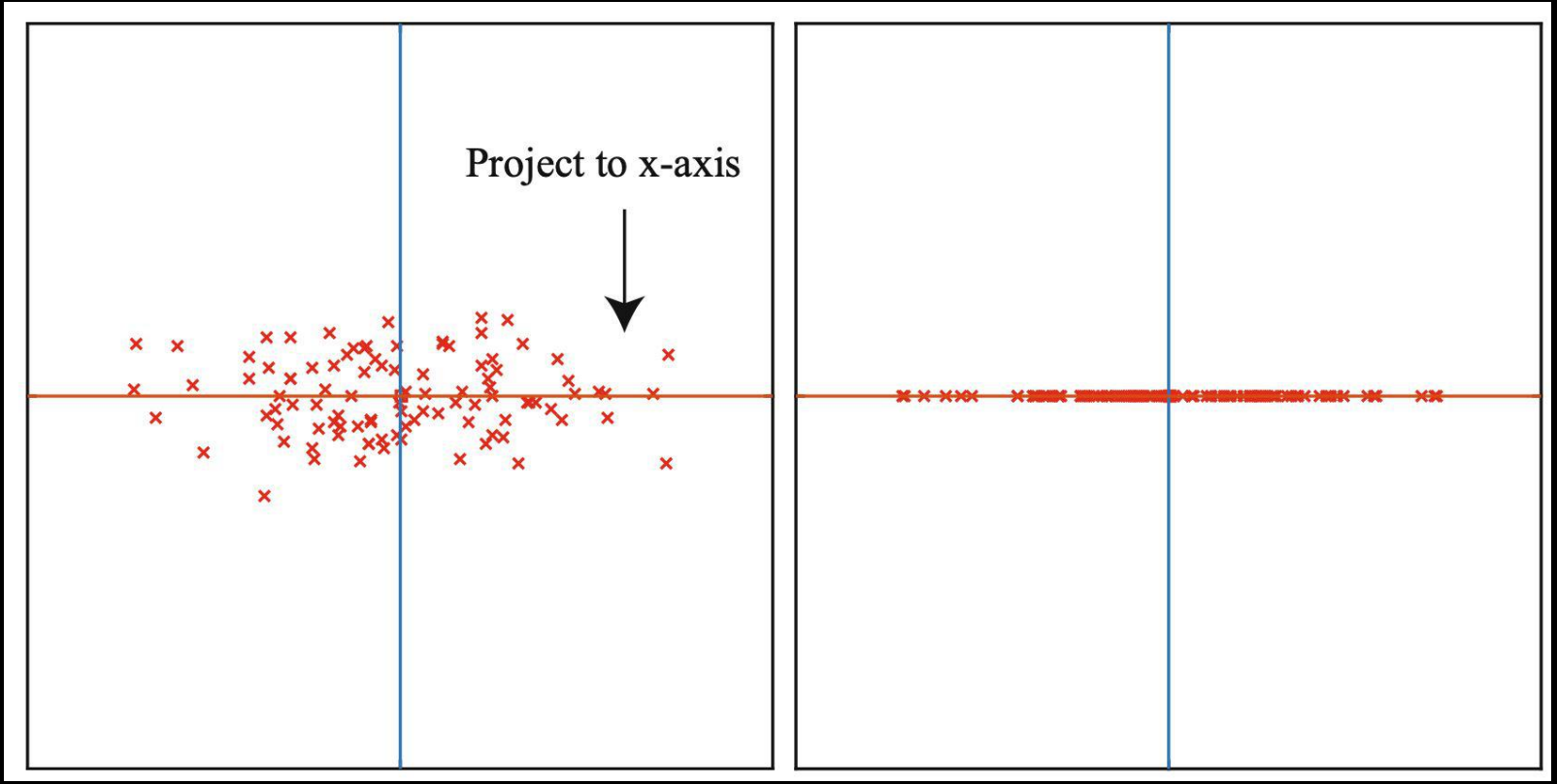
$$\begin{aligned}\text{Covmat}\left(\{\mathbf{m}\}\right) &= \text{Covmat}\left(\{\mathcal{A}\mathbf{x}+\mathbf{b}\}\right)\\ &= \frac{\sum_i(\mathbf{m}_i-\text{mean}\left(\{\mathbf{m}\}\right))(\mathbf{m}_i-\text{mean}\left(\{\mathbf{m}\}\right))^T}{N}\\ &= \frac{\sum_i(\mathcal{A}\mathbf{x}_i+\mathbf{b}-\mathcal{A}\text{mean}\left(\{\mathbf{x}\}\right)-\mathbf{b})(\mathcal{A}\mathbf{x}_i+\mathbf{b}-\mathcal{A}\text{mean}\left(\{\mathbf{x}\}\right)-\mathbf{b})^T}{N}\\ &= \frac{\mathcal{A}\left[\sum_i(\mathbf{x}_i-\text{mean}\left(\{\mathbf{x}\}\right))(\mathbf{x}_i-\text{mean}\left(\{\mathbf{x}\}\right))^T\right]\mathcal{A}^T}{N}\\ &= \mathcal{A}\text{Covmat}\left(\{\mathbf{x}\}\right)\mathcal{A}^T.\end{aligned}$$

Principal Component Analysis

Rotated data preserves angles, translations preserve distances



Projections on axes

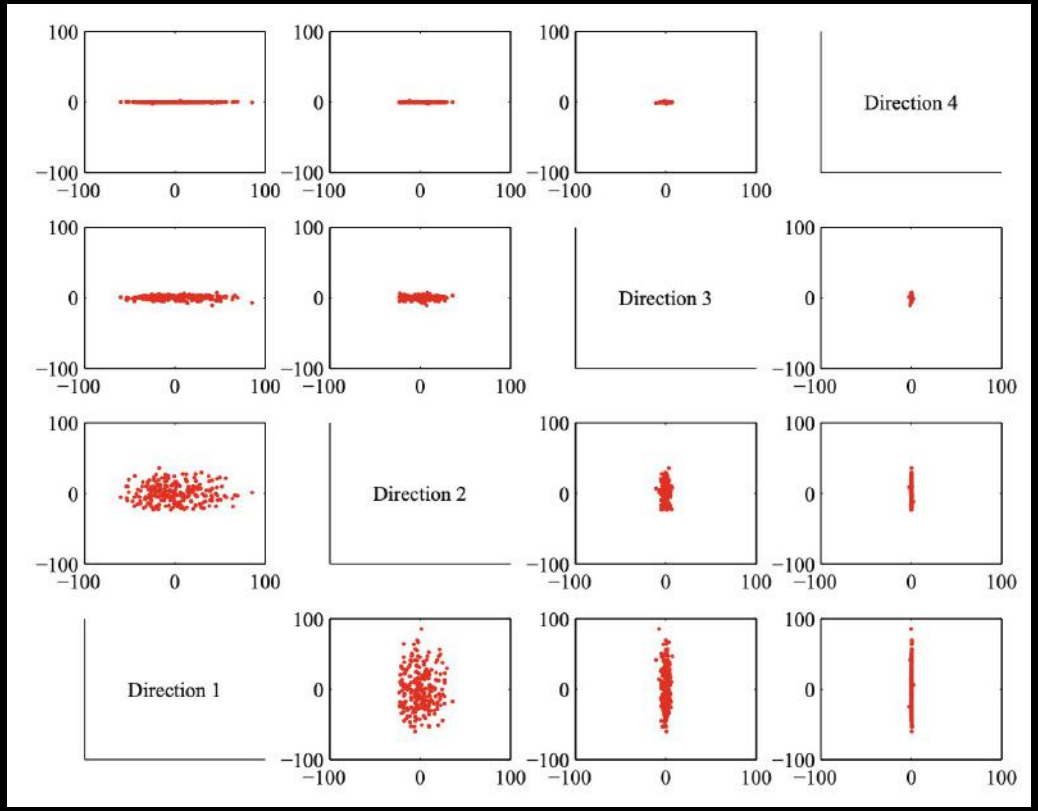


Singular Value Decomposition

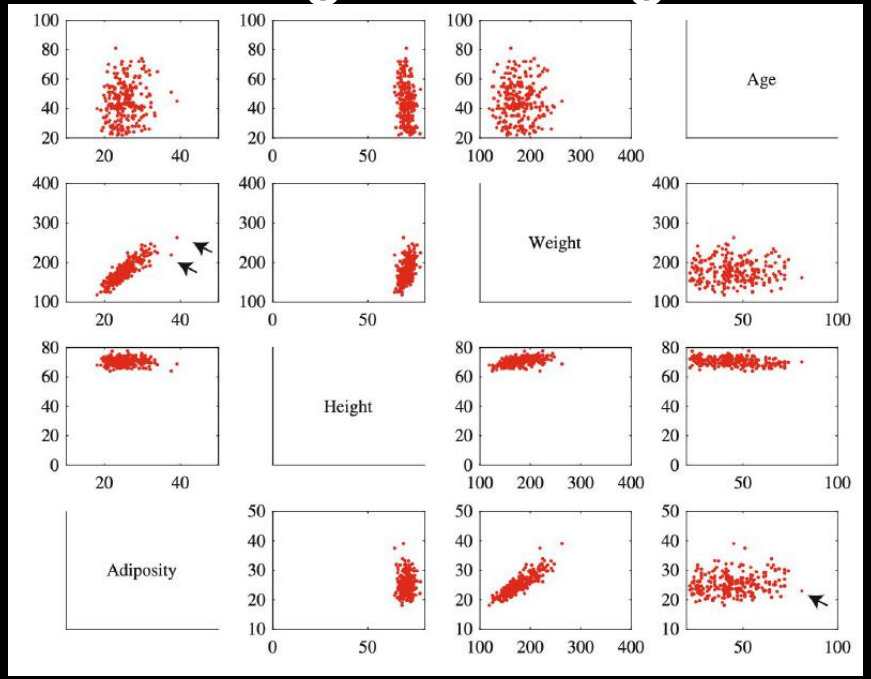
For any $m \times p$ matrix $\mathcal{X}$, it is possible to obtain a decomposition

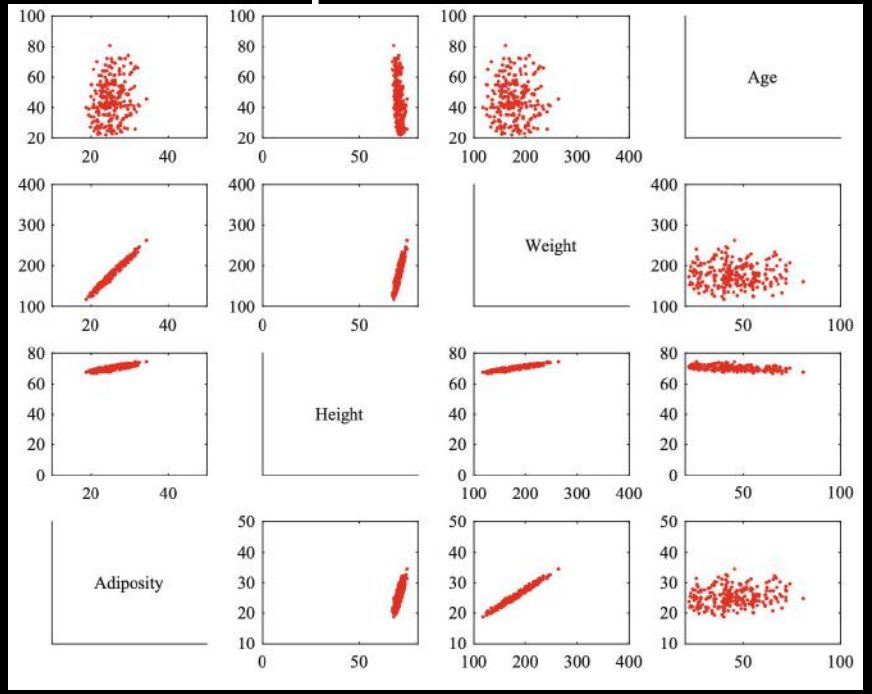
$$\mathcal{X} = \mathcal{U}\Sigma\mathcal{V}^T$$

Principal Components are abstract representations

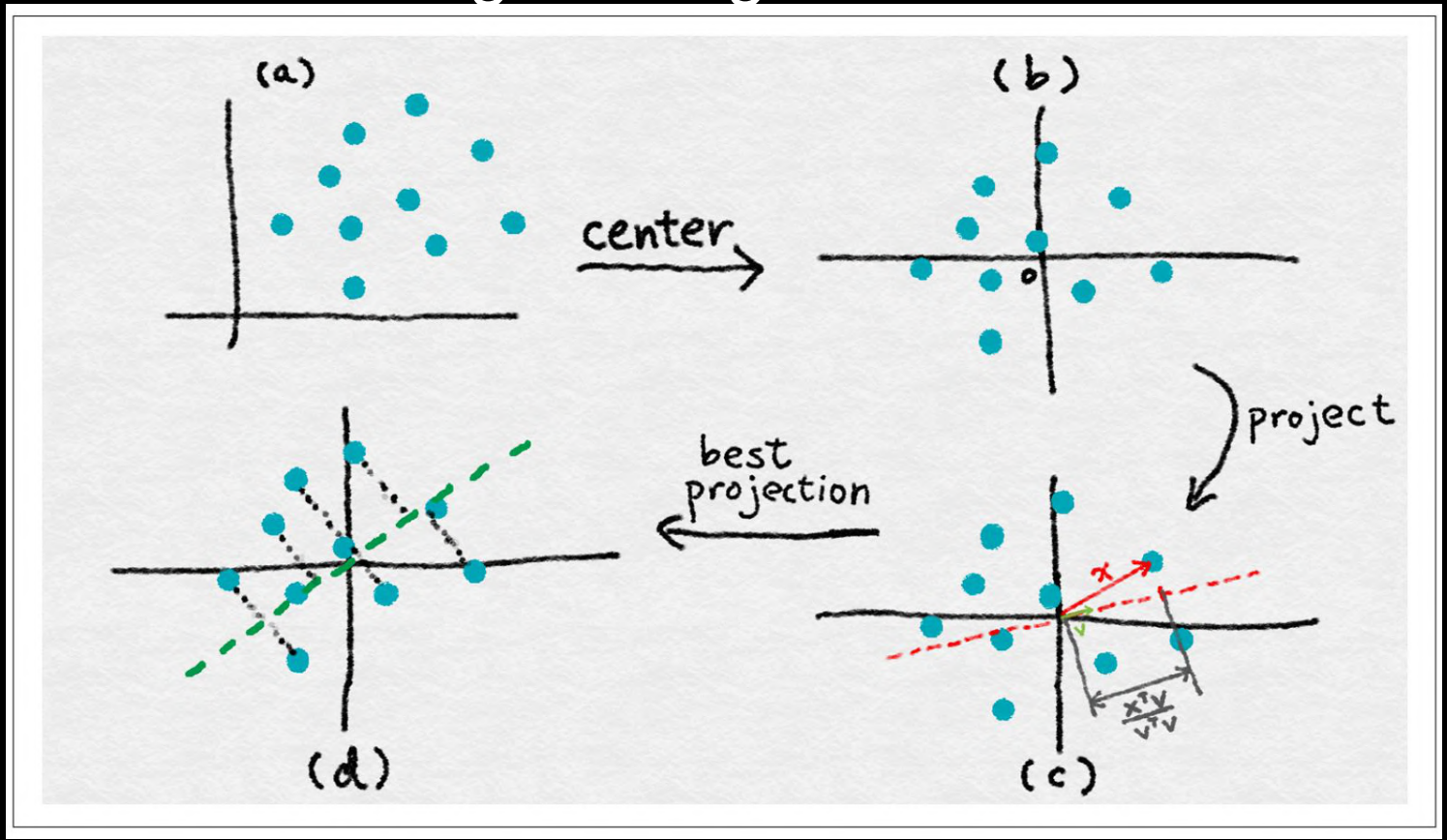


Retrieving back original data from representations

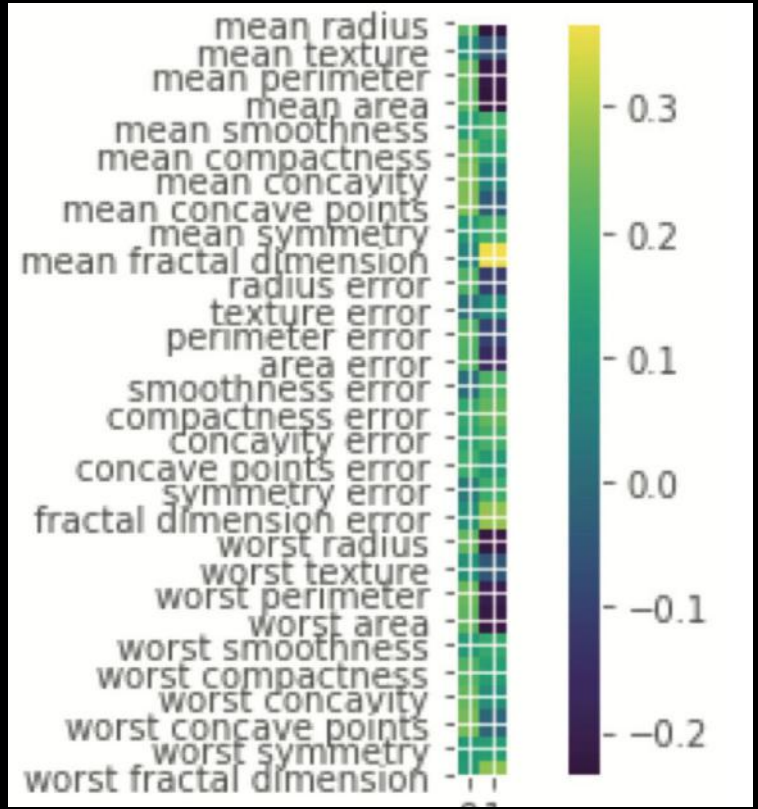




PCA for feature engineering



PCA for feature selection



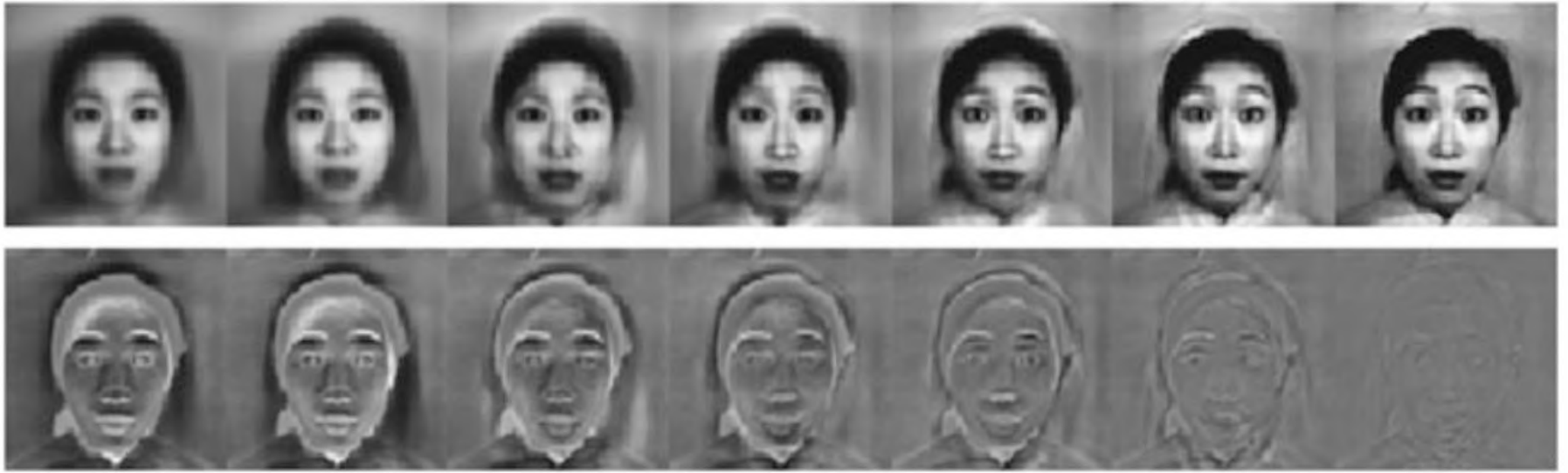
Example: Approximating Faces

Mean image from Japanese Facial Expression dataset

First sixteen principal components of the Japanese Facial Expression dat

Sample Face Image

mean 1 5 10 20 50 100



Approaches for High Dimensional Data

Low rank approximations

- Principal Components Analysis
- Multidimensional scaling
- Factor analysis
- Principal Coordinate Analysis
- Latent Semantic Analysis
- TF-IDF

Machine Learning Approaches

- Filter Based Feature Selection
- Wrapper Based Feature Selection
- Embedded approach
- Bayesian Networks