CAP 5768: Introduction to Data Science

Giri NARASIMHAN

www.cis.fiu.edu/~giri/teach/5768.html

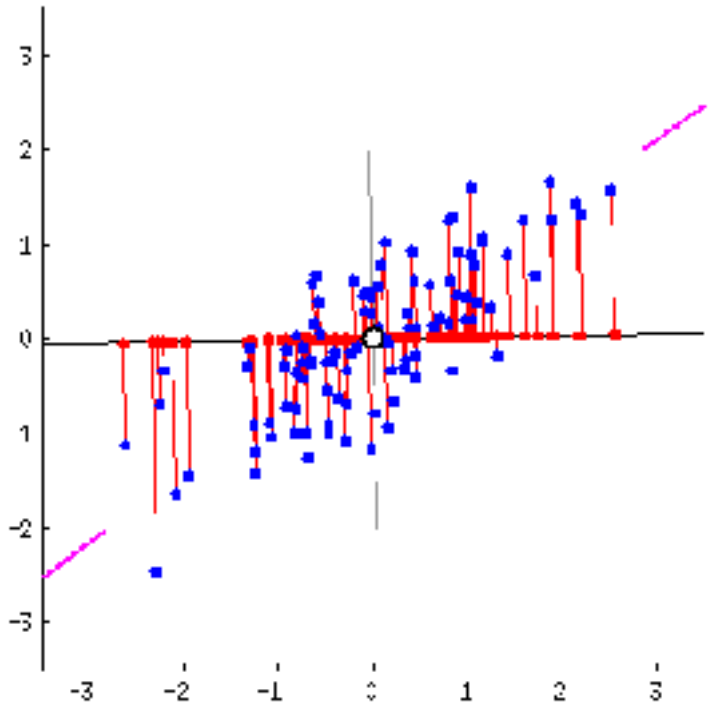
PCA Recap

From **Johnson & Wichern**, ***Applied multivariate statistical analysis***, 6th Ed

PCA

- Tool for Dimensionality Reduction
  - Reduces impact of curse of dimensionality
- Tool for finding Subspace in which data lies
- Summarization of data to find important variables

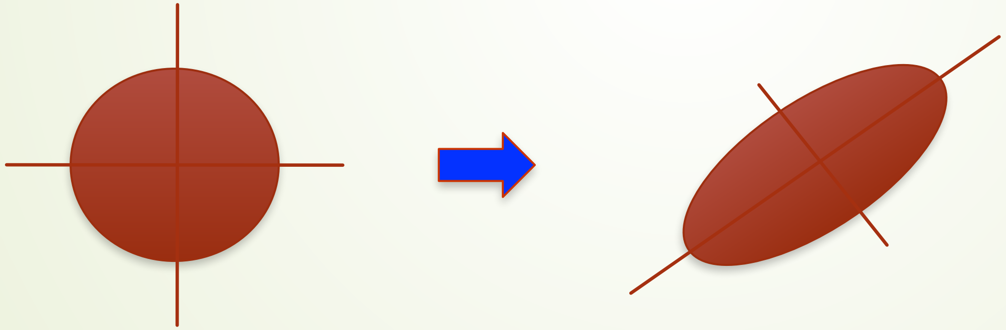
PCA Animation



https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

Matrices & Transformations

- Linear Transformations
  - Ax = y

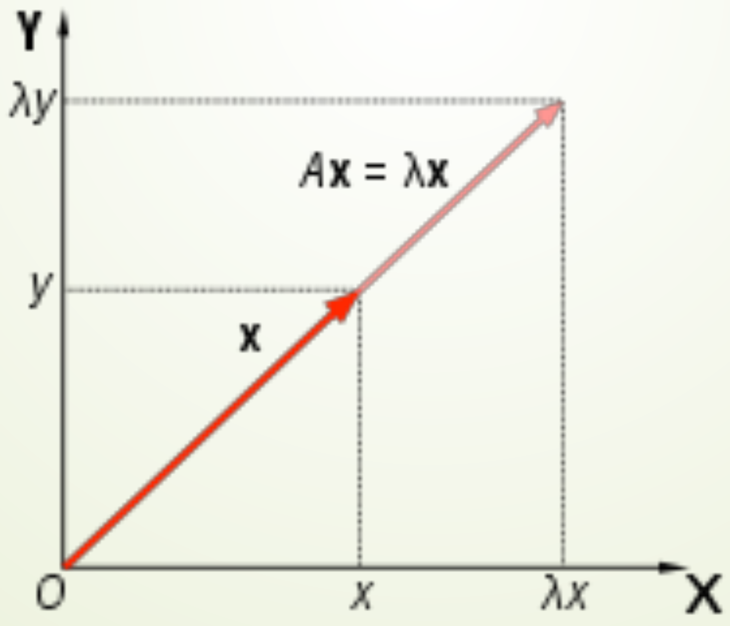


Data as Matrices

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} \begin{array}{l} \leftarrow \text{1st (multivariate) observation} \\ \\ \\ \leftarrow n\text{th (multivariate) observation} \end{array}$$

Eigenvalues and Eigenvectors

- $Ax = \lambda x$, for square matrices A
- Characteristic Eq: $|A - \lambda I| = 0$



Quadratic Form

- The scalar x'Ax is called **quadratic form**

$$Q(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x},$$

$$Q(\mathbf{x}) = \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}x_i x_j.$$

Spectral Decomposition

- For symmetric square matrices A, the spectral decomposition is:

$$\underset{(k\times k)}{\mathbf{A}} = \lambda_1 \underset{(k\times 1)}{\mathbf{e}_1} \underset{(1\times k)}{\mathbf{e}_1'} + \lambda_2 \underset{(k\times 1)}{\mathbf{e}_2} \underset{(1\times k)}{\mathbf{e}_2'} + \cdots + \lambda_k \underset{(k\times 1)}{\mathbf{e}_k} \underset{(1\times k)}{\mathbf{e}_k'}$$

$$\boxed{\mathbf{A} = \sum_{i=1}^{k} \lambda_i \mathbf{e}_i \mathbf{e}_i'}$$

Spectral Decomposition … 2

$$\underset{(k\times k)}{\mathbf{A}} = \sum_{i=1}^{k} \lambda_i \underset{(k\times 1)}{\mathbf{e}_i} \underset{(1\times k)}{\mathbf{e}_i'} = \underset{(k\times k)}{\mathbf{P}} \underset{(k\times k)}{\mathbf{\Lambda}} \underset{(k\times k)}{\mathbf{P}'}$$

$$\mathbf{P} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k]$$

$$\underset{(k\times k)}{\mathbf{\Lambda}} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix}$$

Dimension Reduction Revisited

- If we take r eigenvectors, then
  - $P_r = [e_1, e_2, \ldots, e_r]$, and

$$\underset{(r \times r)}{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_r \end{bmatrix}$$

  - A can be approximated by taking r eigenvectors

$$\underset{(k \times r)}{P} \; \underset{(r \times r)}{\Lambda} \; \underset{(r \times k)}{P'}$$

Singular Value Decomposition

- Spectral Decomp. for sq. symm. matrices
- Non-sq. asymmetric matrices?
  - Use sq. root of eigenvalues of AA’
  - Singular values of A

$$\mathbf{P} \quad \mathbf{\Lambda} \quad \mathbf{P'}$$

(k X r) (r X r) (r X k)

$$\underset{(m\times k)}{\mathbf{A}} = \underset{(m\times m)}{\mathbf{U}} \; \underset{(m\times k)}{\mathbf{\Lambda}} \; \underset{(k\times k)}{\mathbf{V'}}$$

Dimensionality Reduction

- Given m X k matrix A, we can approximate it by m X s matrix B with s < k = rank(A). Then

$$\mathbf{B} = \sum_{i=1}^{s} \lambda_i \mathbf{u}_i \mathbf{v}_i'$$

- Here we are picking s singular values from SVD

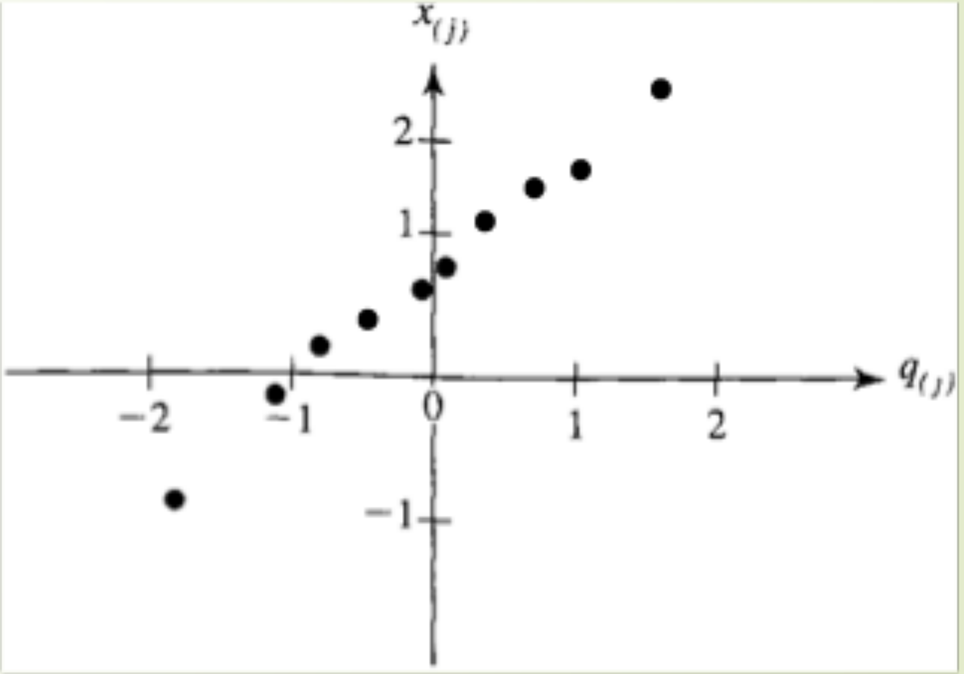
Central Limit Theorem

How to make data “normal”?

- Let X_1, X_2, …, X_n be independent observations from any distribution with mean μ and variance ∑. Then
  - $\sqrt{n}\,(\overline{\mathbf{X}} - \boldsymbol{\mu})$ has an approximate $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ distribution
  - , $\mathbf{X}_n$ be independent observations from a populat
- Sample size, n, must be large relative to p

Q-Q plot

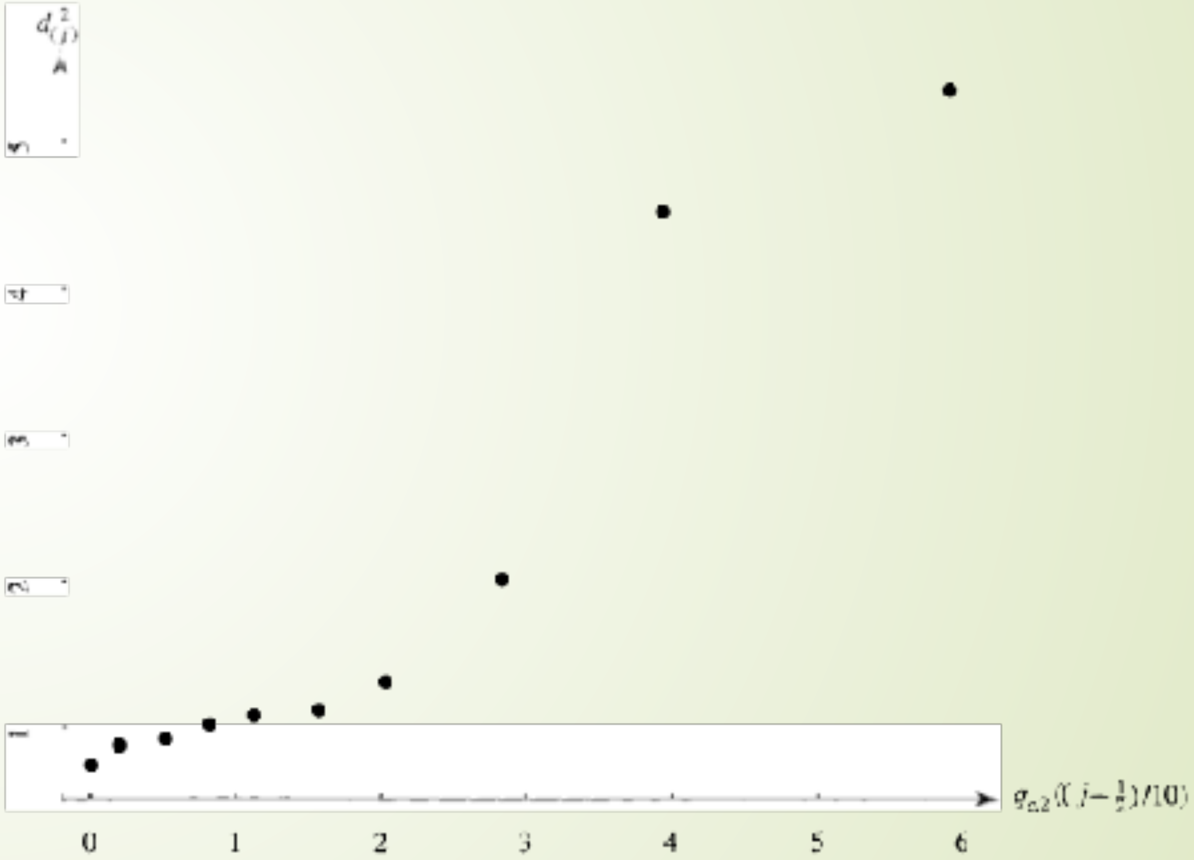
Ordered observations $x_{(j)}$	Probability levels $\left(j - \frac{1}{2}\right)/n$	Standard normal quantiles $q_{(j)}$
−1.00	.05	−1.645
−.10	.15	−1.036
.16	.25	−.674
.41	.35	−.385
.62	.45	−.125
.80	.55	.125
1.26	.65	.385
1.54	.75	.674
1.71	.85	1.036
2.30	.95	1.645



Chi-Square Plots

$$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}), \qquad j = 1, 2, \ldots, n$$

j	$d_{(j)}^2$	$q_{c,2}\left(\frac{j-\frac{1}{2}}{10}\right)$
1	.59	.10
2	.81	.33
3	.83	.58
4	.97	.86
5	1.01	1.20
6	1.02	1.60
7	1.20	2.10
8	1.88	2.77
9	4.34	3.79
10	5.33	5.99



Chi-Square DIstribution

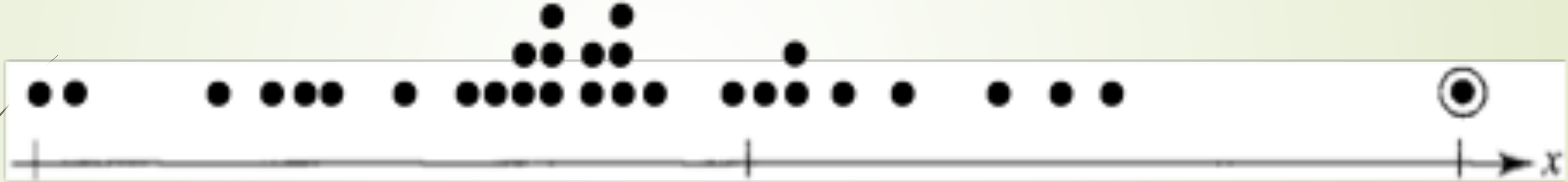
- Squared **Generalized Distances**

$$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}), \qquad j = 1, 2, \ldots, n$$

- If X is multivariate normal and n and n-p are large, then the squared distances behave like a **chi-squared** plot or **gamma** plot.

Detecting Outliers

- Visual detection



- Harder in multivariate case. Why?
  - May be univariate or multivariate outlier

Bivariate Outliers

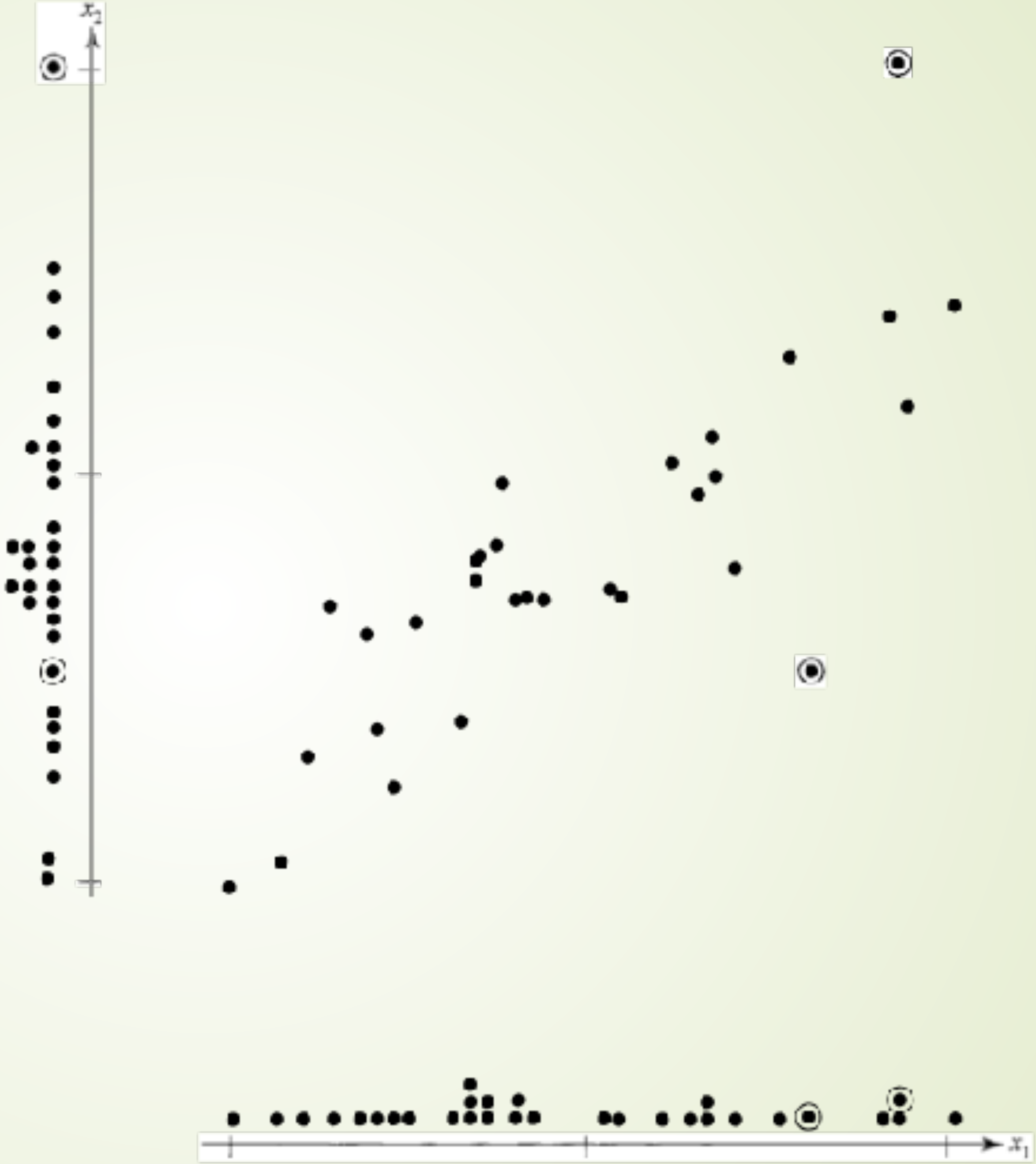


Figure 4.10 Two outliers; one univariate and one bivariate.

Multivariate Outliers

- Some outliers are hard to detect
- Look for large values of
  - $(\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}).$

Outlier detection

- Dot plots for each variable
- Scatter plot for each pair of variables
- Calculate z-values and examine for outliers

$$z_{jk} = (x_{jk} - \bar{x}_k)/\sqrt{s_{kk}}$$

- Calculate gen sq distances & look for outliers

$$(\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}).$$

Other Transforms for Normality

HELPFUL TRANSFORMATIONS TO NEAR NORMALITY

Original Scale	*Transformed Scale*	
1. Counts, y	$\sqrt{y}$	
2. Proportions, $\hat{p}$	$\text{logit}(\hat{p}) = \frac{1}{2}\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$	(4-33)
3. Correlations, r	Fisher's $z(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$	