Principal Component Analysis

Goals

- **Restructure interrelated variables**
- **Simplify description**
- **Reduce Dimensionality**
- **Avoid multi-collinearity problems in regression**

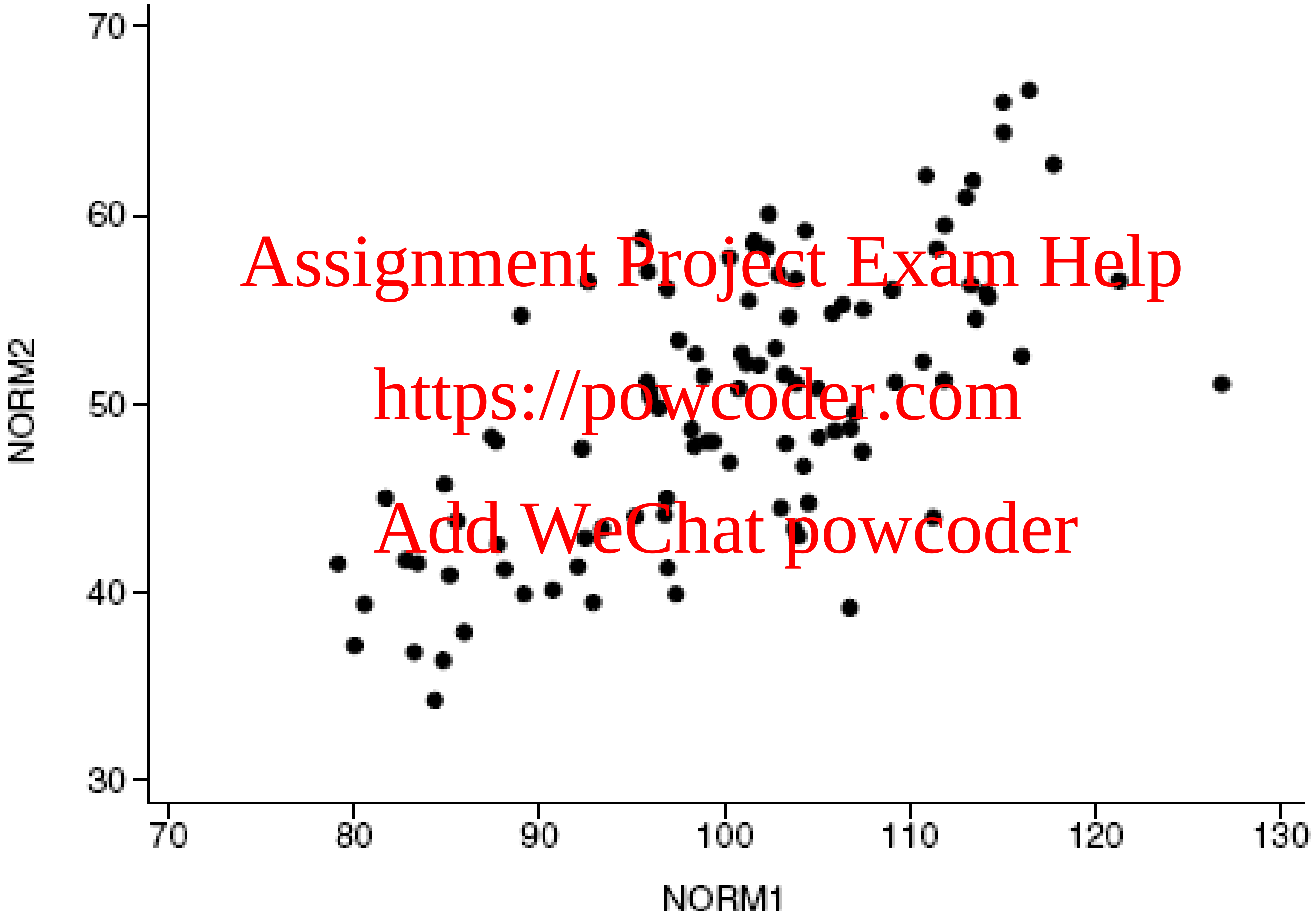
Basic Idea

- X_1 , X_2 – correlated

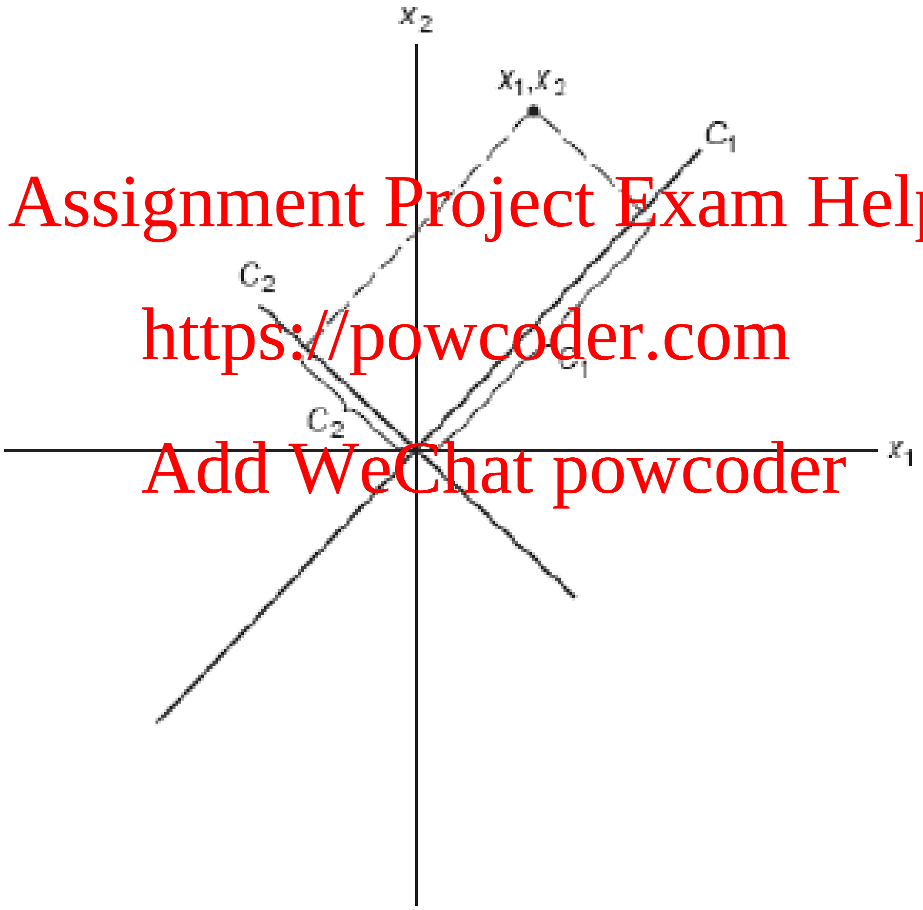
Transform them into:

- C_1 , C_2 – uncorrelated

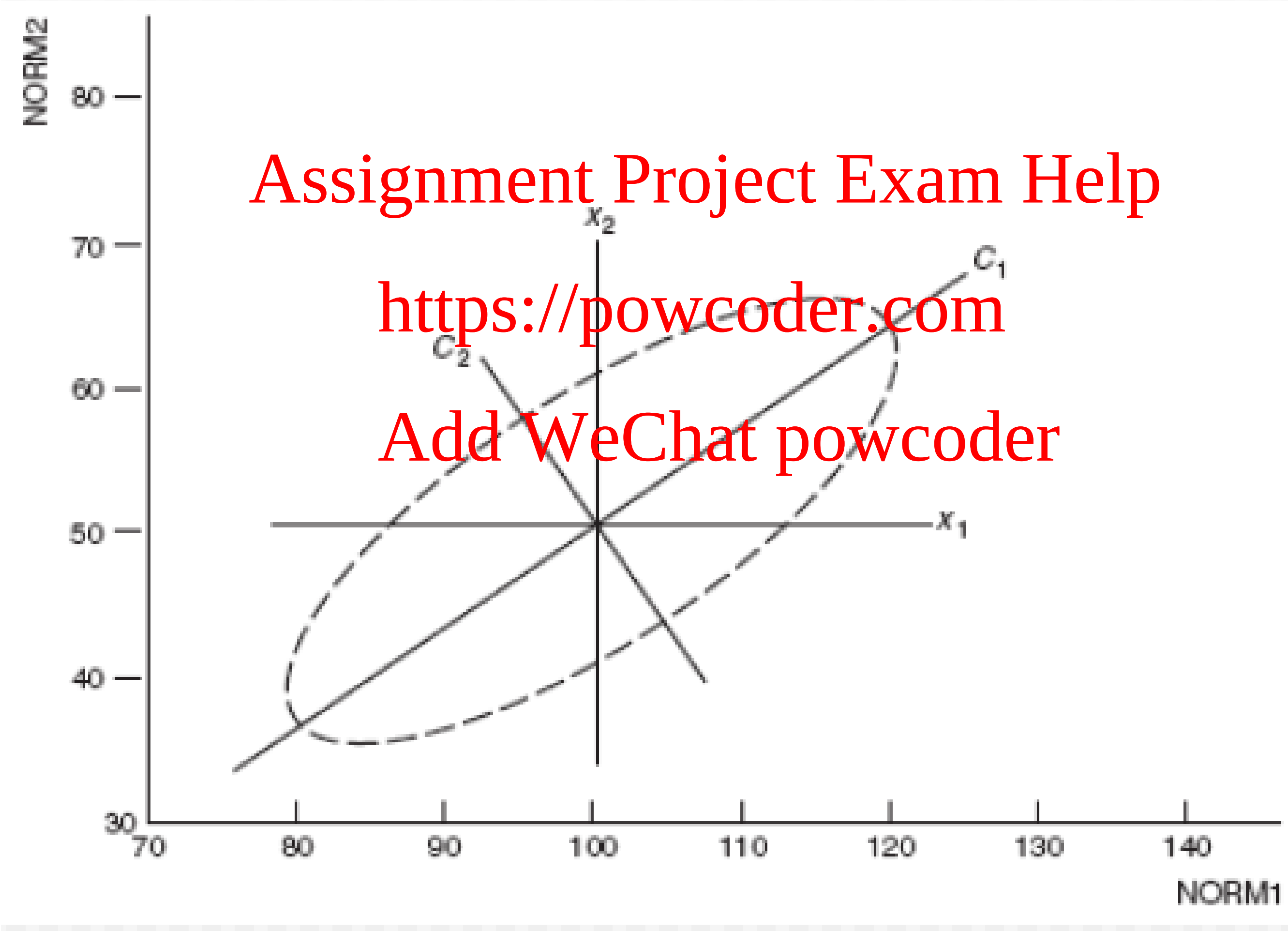
Geometric Concept



2 Principal Components: C_1 , C_2 (p 361)



Plot of Principal Components: C_1 , C_2 (p 362)



Equations for Principal Components: C_1 , C_2

- $C_1 = 0.85X_1 + 0.53X_2$
- $C_2 = -0.53X_1 + 0.85X_2$
- Var $C_1 = 135.0$
- Var $C_2 = 22.5$

- Note: Dot Product of C_1 , C_2 = 0
  - Hence they are orthogonal, therefore uncorrelated, and Var $C_1 >$ Var C_2

Principal Component Model

Matrix Equation

$$\begin{vmatrix} C_1 \\ C_2 \\ \vdots \\ C_p \end{vmatrix} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ & & \vdots & \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{vmatrix} \begin{vmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{vmatrix}$$

Principal Components: C_1 ,…, C_p

- **From P original variables: X_1 ,…, X_p derive P principal components C_1 ,…, C_p**
- **Each C_j is a linear combination of the X_i 's**
- **$C_j = a_{j1} X_1 + a_{j2} X_2 + \ldots + a_{jp} X_p$**

Properties of Principal Components:

- **Coefficients are chosen to satisfy:**

$$\mathbf{Var\ C_1 \geq Var\ C_2 \geq \ . \ \ . \ \ . \geq Var\ C_p}$$

- **Variance is a measure of information:**
  - **For Example, in Prostate Cancer:**

- **Any two principal components are orthogonal, hence uncorrelated**

Calculation of Principal Components

- **Let S be the Covariance Matrix of the X variables.**

$$S = \begin{vmatrix} s_{11} & s_{12} & \ldots & s_{1p} \\ s_{21} & s_{22} & \ldots & s_{2p} \\ & & \vdots & \\ s_{p1} & s_{p2} & \ldots & s_{pp} \end{vmatrix}$$

- **Then a_{ij} 's are the solution to the equation:**

$$(S - \lambda I)a = 0$$

(Hotelling – 1933)

Recall Some Terminology

- Solutions to **(S – λI)a = 0** are:
  - **λ** a scalar known as the eigenvalue
  - **a** a vector known as the eigenvector
- **a** is not unique. There are an infinite number of possibilities, so:
  - Choose **a** such that the sum of the squares of coefficients for each eigenvector is = 1.
  - This yields: P unique eigenvalues and P corresponding eigenvectors.

Then

- The eigenvectors are the Principal Components
  - **λ** a scalar known as the eigenvalue
  - **a** a vector known as the eigenvector

- **a** is not unique. There are an infinite number of possibilities, so:

  - Choose **a** such that the sum of the squares of coefficients for each eigenvector is = 1.

  - This yields: P unique eigenvalues and P corresponding eigenvectors.

Review

- Principal Components are the **eigenvectors,**
  - and their variances are the **eigenvalues**

  - of the covariance matrix **S** of the **X**'s

- Variances of the $\mathbf{C}_i$ **'**s add to the sum of the

  variances of the original variables (total variances)

Example Revisited (p 361)

- **Var X_1 = 104.0 Var X_2 = 53.5 sum = 157.5**
- **Var C_1 = 135.0 Var C_2 = 22.5 sum = 157.5**
- **Total Variance is Preserved**

Choosing **m**

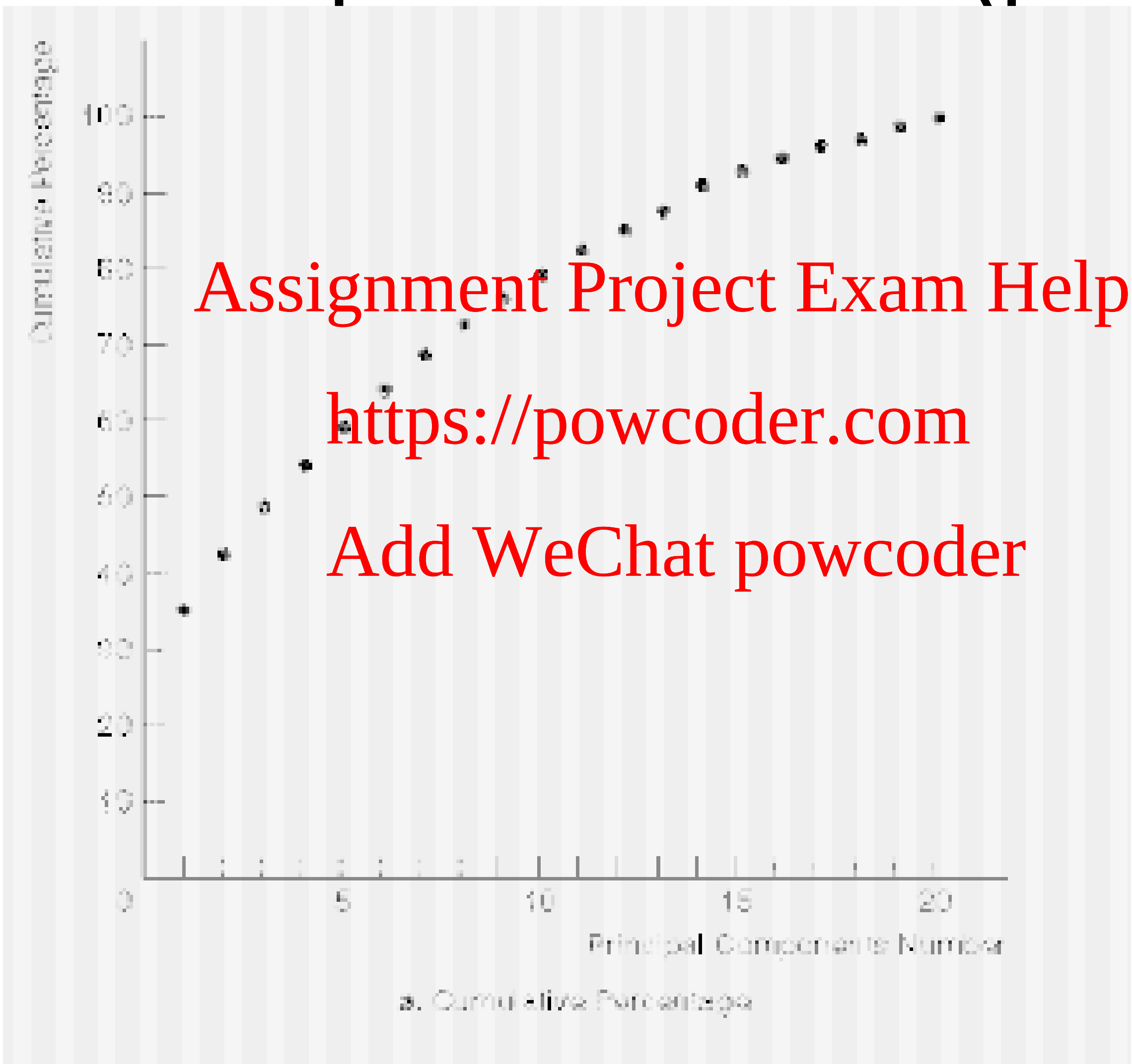
- **Rely on existing theory**
- **Kaiser's Rule:**

  - **S: choose λ_i > sum of variance of X's/P**

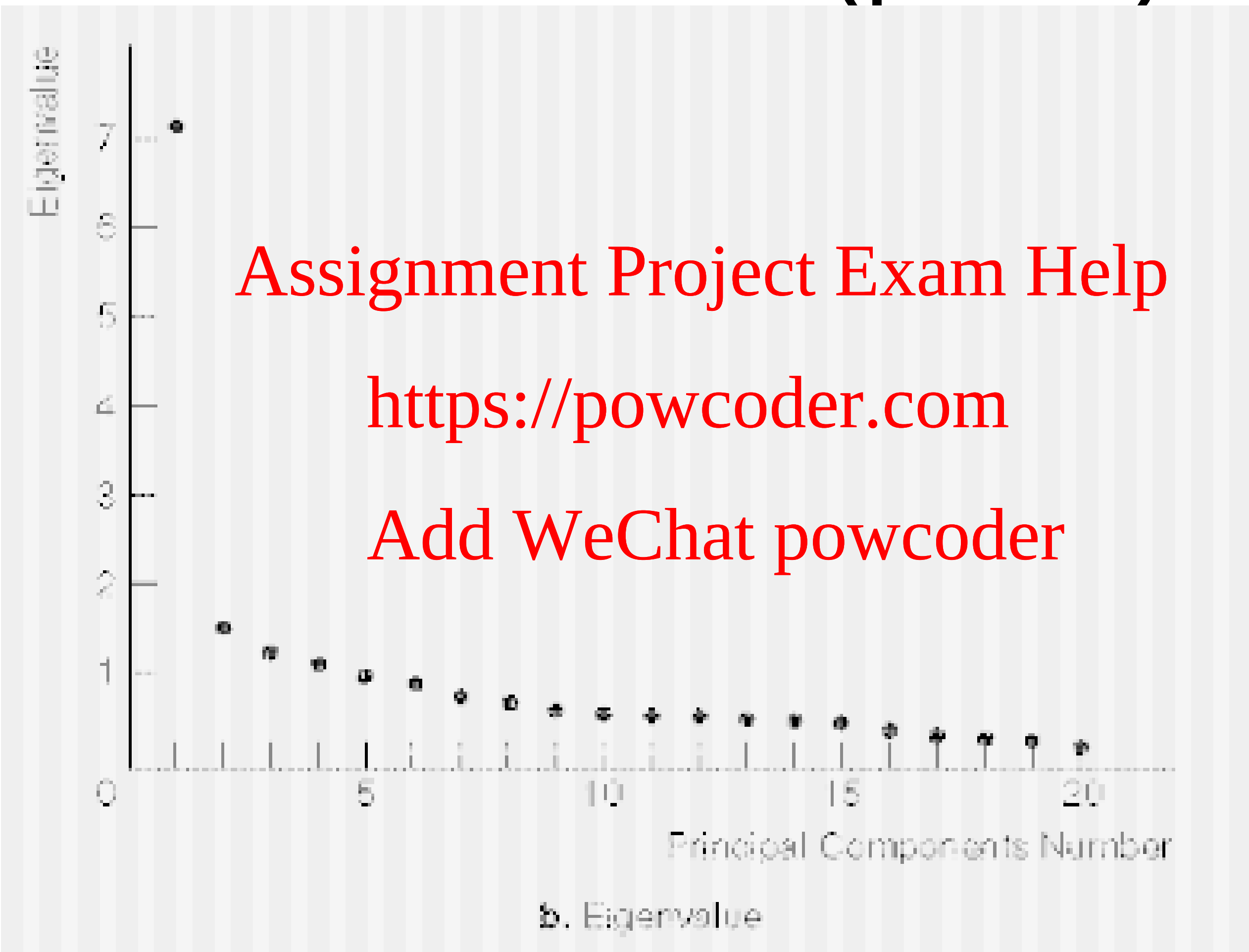
  - **R: choose λ_i > 1**

- **We need to explain a given %**
- **Elbow Rule**

Cumulative Percentages of Total Variance for Depression Data (p 368)



Eigenvalues for Depression Data – Scree Plot (p 368)



A plot, in descending order of magnitude, of the eigenvalues of a correlation matrix

Depression, CESD, Example

- **Using R (Correlation, Not Language):**
  - **Choose Eigenvalues $>$ 1**

- **Hence choose 5 Principal Components**

Elbow Rule for Choosing m

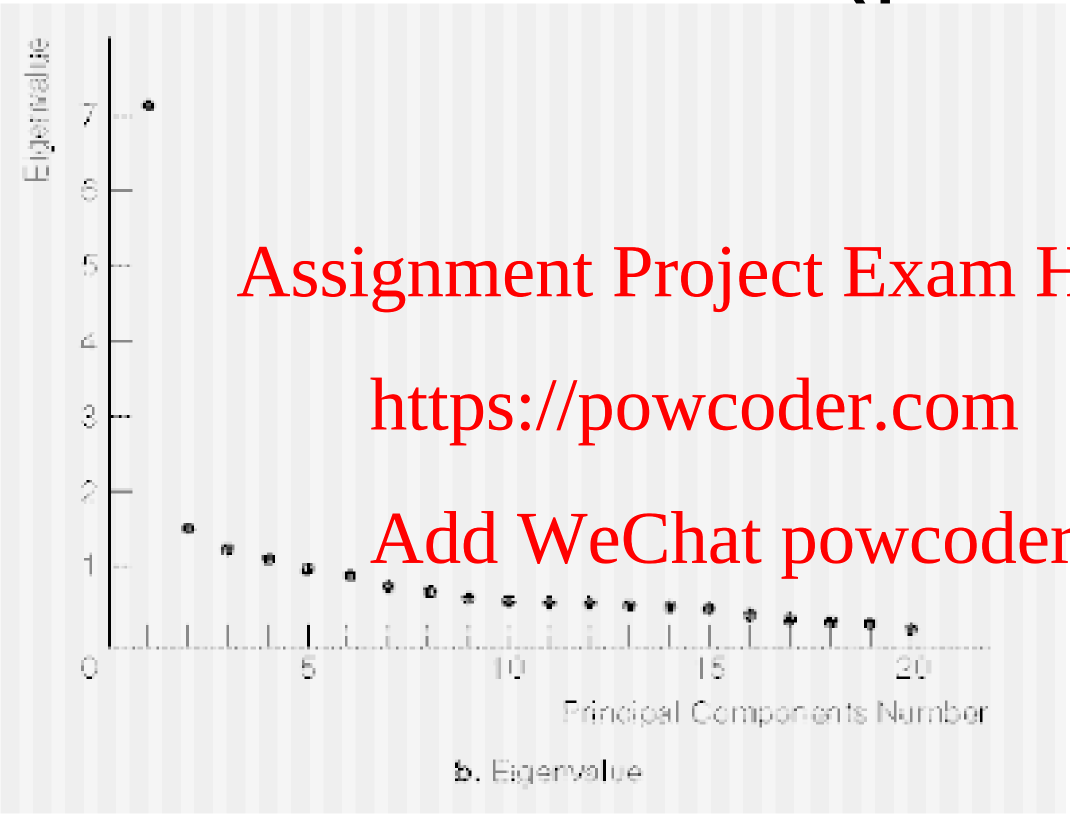
- **Start with the scree plot**
- **Choose a cutoff point where:**

  - **Lines joining consecutive points are “steep” left of the cutoff point, and**

  - **“flat” right of the cutoff point**

- **The point where the two slopes meet is the cutoff point**

Eigenvalues for Depression Data – Scree Plot (p 368)



<u>M = 2</u>

Principal Components for standardized CESD scale items (p 369)

Table 14.2: Principal components analysis for standardized CESD scale items (depression data set)

Item	Principal component 1	2	3	4	5
Negative affect					
I felt that I could not shake off the blues even with the help of my family or friends.	0.2774	[illegible]	[illegible]	[illegible]	[illegible]
I felt depressed.	0.3139	−0.0971	0.0316	0.2475	0.0944
I felt lonely.	0.2679	0.1547	0.0346	0.2472	−0.2183
I had crying spells.	0.2436	0.3194	0.1769	−0.0716	−0.1729
I felt sad.	0.2868	0.0497	0.1384	0.2794	−0.0411
I felt fearful.	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
I thought my life had been a failure.	0.2914	0.1621	−0.0190	−0.0761	−0.0870
Positive affect					
I felt that I was as good as other people.	0.1061	0.3045	0.1163	−0.5567	−0.0776
I felt hopeful about the future.	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
I was happy.	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
I enjoyed life.	0.2135	0.1048	−0.3314	0.0414	0.2419
Somatic, retarded activity					
I was bothered by things that usually don't bother me.	0.1796	0.2165	0.1534	0.1451	0.0368
I did not feel like eating; my appetite was poor.	0.1259	−0.2125	0.2645	−0.5406	0.0953
I felt that everything was an effort.	0.1803	0.4613	0.1914	0.2461	0.0847
My sleep was restless.	0.2004	−0.2098	0.2703	0.0312	0.0834
I could not 'get going.'	0.1924	−0.4171	−0.1650	−0.0467	−0.0399
I had trouble keeping my mind on what I was doing.	0.2097	−0.3883	−0.0660	−0.0684	−0.0499
I talked less than usual.	0.1717	−0.0163	0.2019	−0.0629	0.2752
Interpersonal					
People were unfriendly.	0.1315	−0.0563	−0.6326	−0.0339	−0.3349
I felt that people disliked me.	0.2357	0.2283	−0.1932	−0.2404	−0.2909
Eigenvalues or Var C_i	7.055	1.486	1.231	1.066	1.013
Cumulative proportion	0.353	0.427	0.489	0.542	0.593
$0.5/(\text{VAR } C_i)^{1/2}$	0.188	0.410	0.451	0.484	0.497

Reading the Output

- **Here:** X_1 **= "I felt that I could not shake …"**

  X_2 **= "I felt depressed, …"**

- **The Principal Component are:**

  $C_1 = 0.2774X_1 + 0.3132X_2 + \ldots$

  $C_2 = 0.1450X_1 + 0.0271X_2 + \ldots$

  etc.

Coefficients as Correlations

- **Recall: Correlation (C_i , X_j) = $a_{ij} * \lambda_i^{½}$**
- **Choose X's where Correlation > 0.5**
  - **Example:**
  - **For C_1 , λ_1 = 7.055; $\lambda_1^{½}$ = 2.656**
  - **Correlation > 0.5 => a_{1j} > 0.5/2.656 = 0.188**
- **Similarly for other Principal Components**