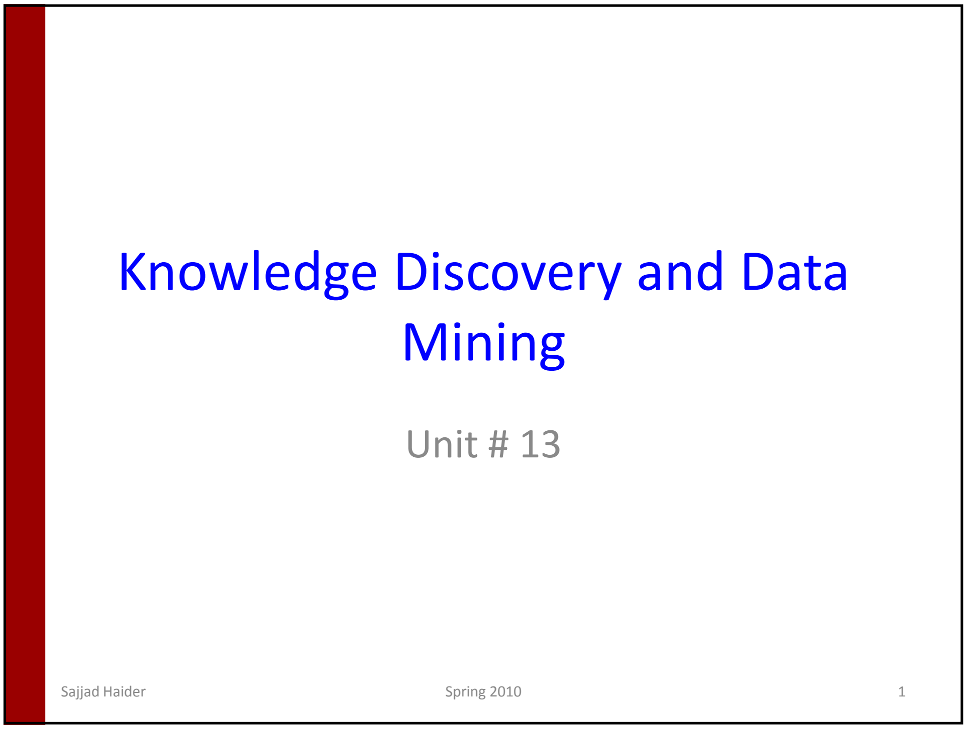
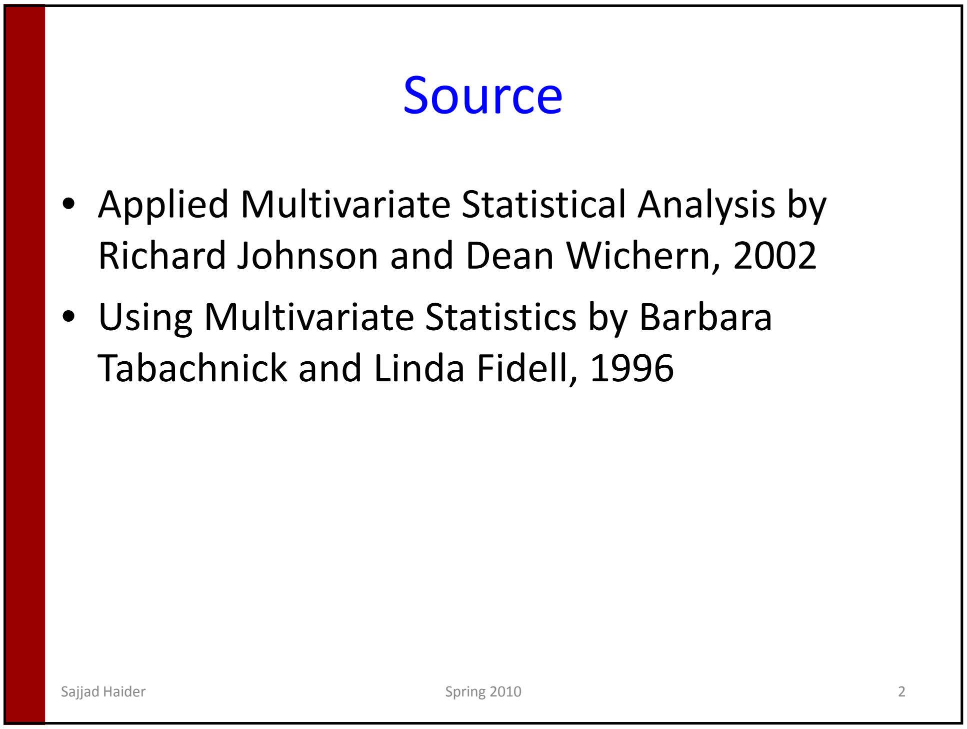


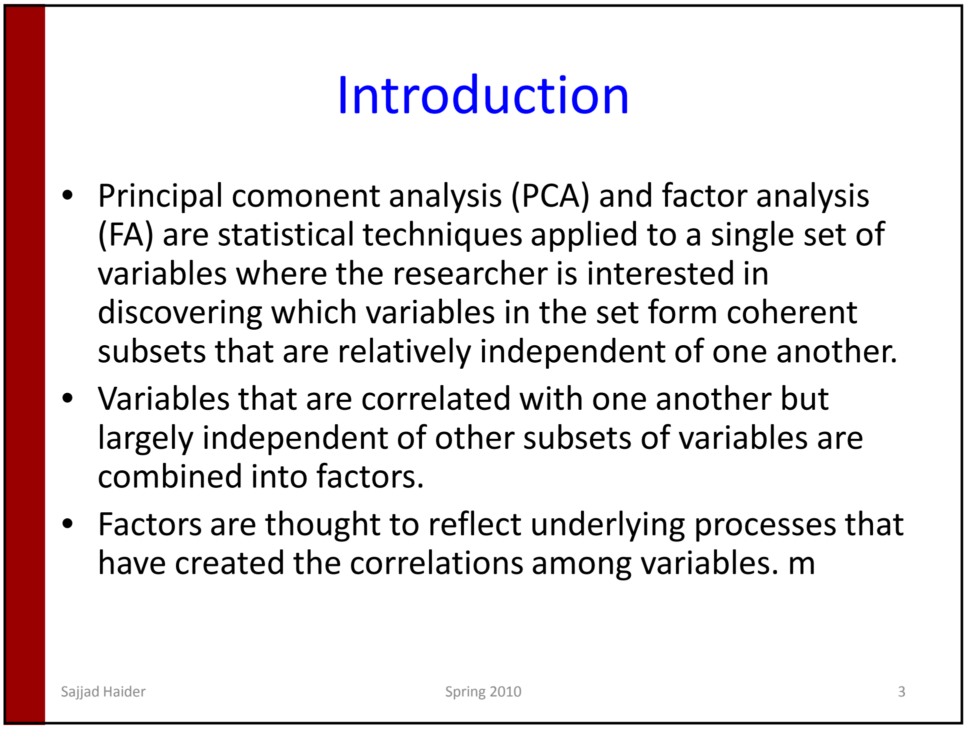
Knowledge Discovery and Data Mining

Unit # 13



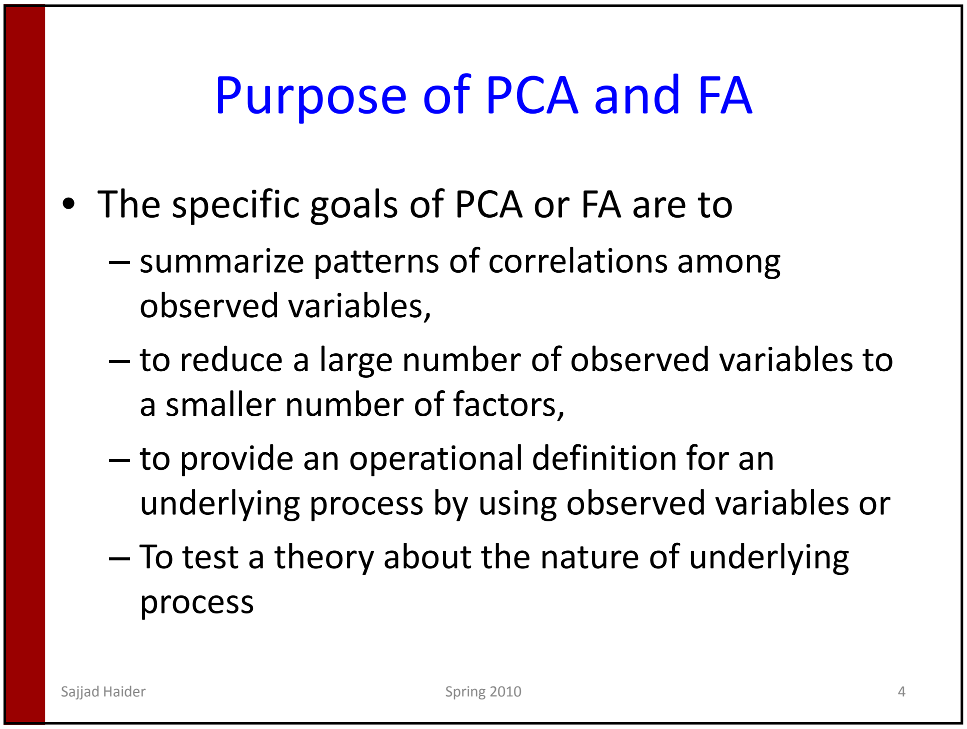
Source

- Applied Multivariate Statistical Analysis by Richard Johnson and Dean Wichern, 2002
- Using Multivariate Statistics by Barbara Tabachnick and Linda Fidell, 1996



Introduction

- Principal comonent analysis (PCA) and factor analysis (FA) are statistical techniques applied to a single set of variables where the researcher is interested in discovering which variables in the set form coherent subsets that are relatively independent of one another.
- Variables that are correlated with one another but largely independent of other subsets of variables are combined into factors.
- Factors are thought to reflect underlying processes that have created the correlations among variables. m



Purpose of PCA and FA

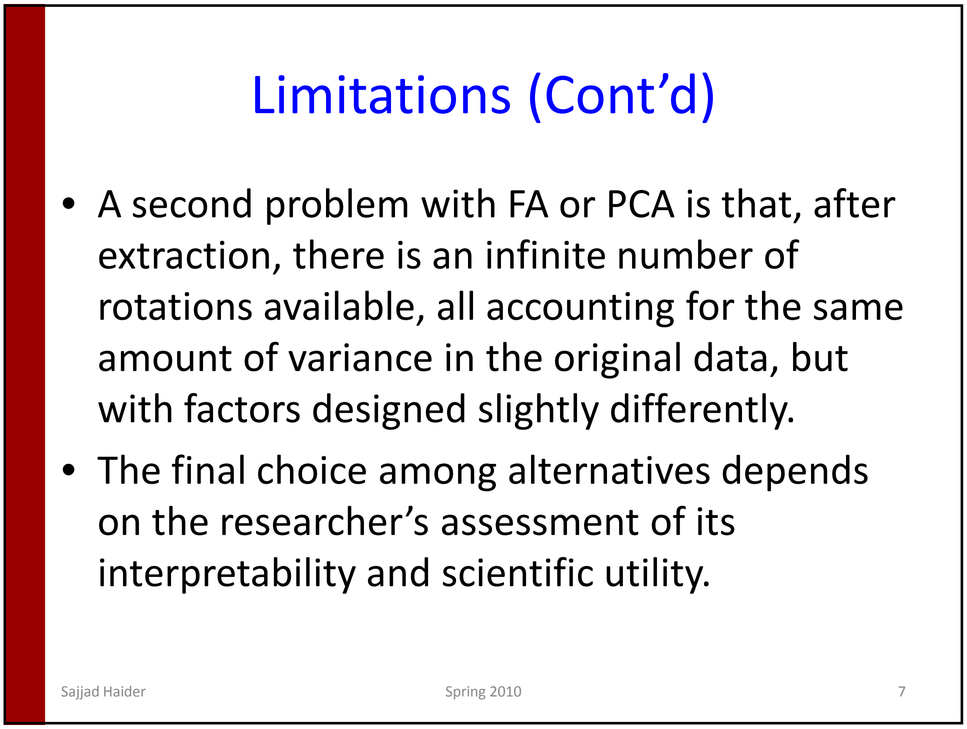
- The specific goals of PCA or FA are to
  - – summarize patterns of correlations among observed variables,
  - – to reduce a large number of observed variables to a smaller number of factors,
  - – to provide an operational definition for an underlying process by using observed variables or
  - – To test a theory about the nature of underlying process

Fundamental Steps

- Steps in PCA or FA include
  - Selecting and measuring a set of variables
  - Preparing the correlation matrix
  - Extracting a set of factors from the correlation matrix
  - Determining the number of factors
  - (probably) rotating the factors to increase interpretability
  - interpreting the results
- Although there are relevant statistical considerations to most of these steps, an important test of the analysis is its interpretability.

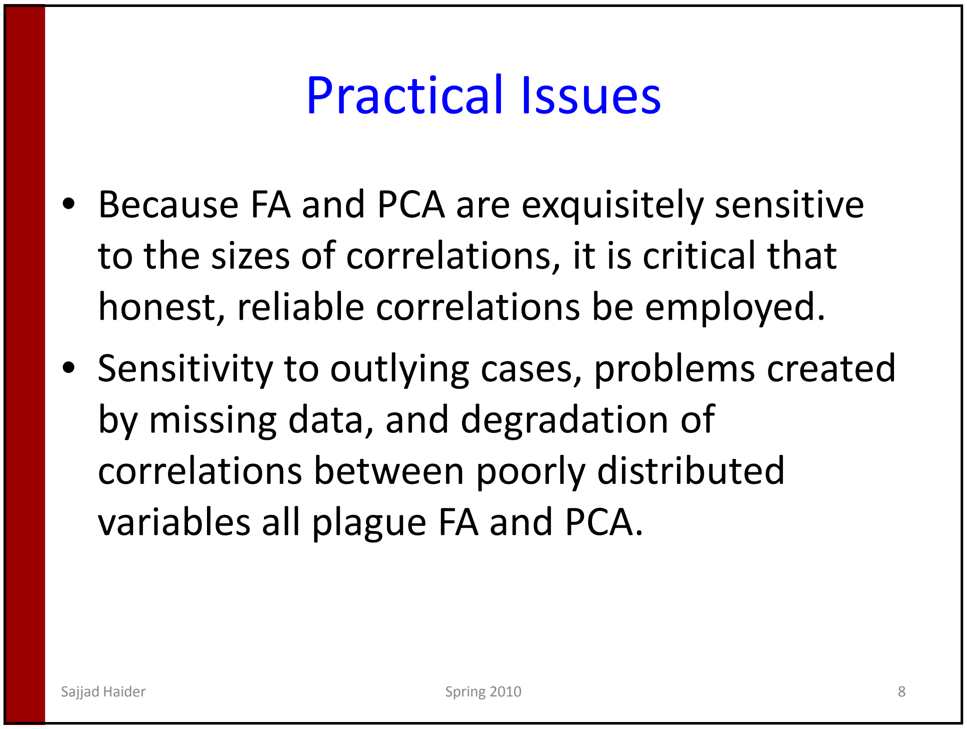
Limitations

- One of the problems with PCA and FA is that there is no criterion variable against which to test the solution.
- In regression analysis, for instance, the dependent variable (DV) is a criterion and the correlation between observed and predicted DV scores serves as a test of the solution
- In classification, the solution is judged by how well it predicts group membership.
- But in PCA or FA there is no external criterion such as group membership against which to test the solution.



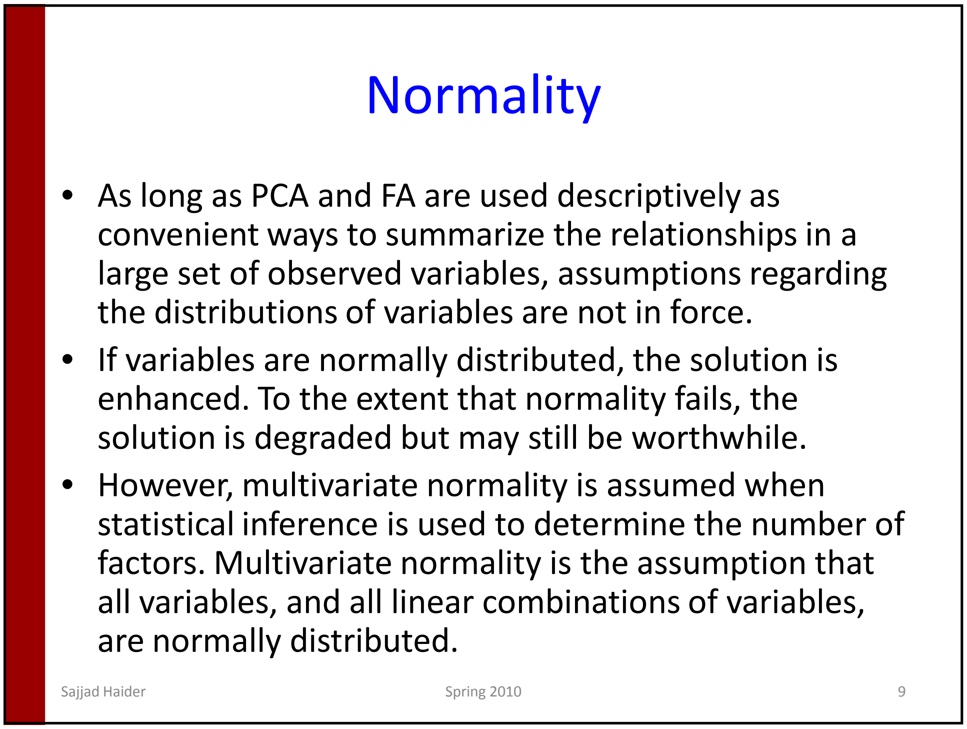
Limitations (Cont’d)

- A second problem with FA or PCA is that, after extraction, there is an infinite number of rotations available, all accounting for the same amount of variance in the original data, but with factors designed slightly differently.
- The final choice among alternatives depends on the researcher’s assessment of its interpretability and scientific utility.



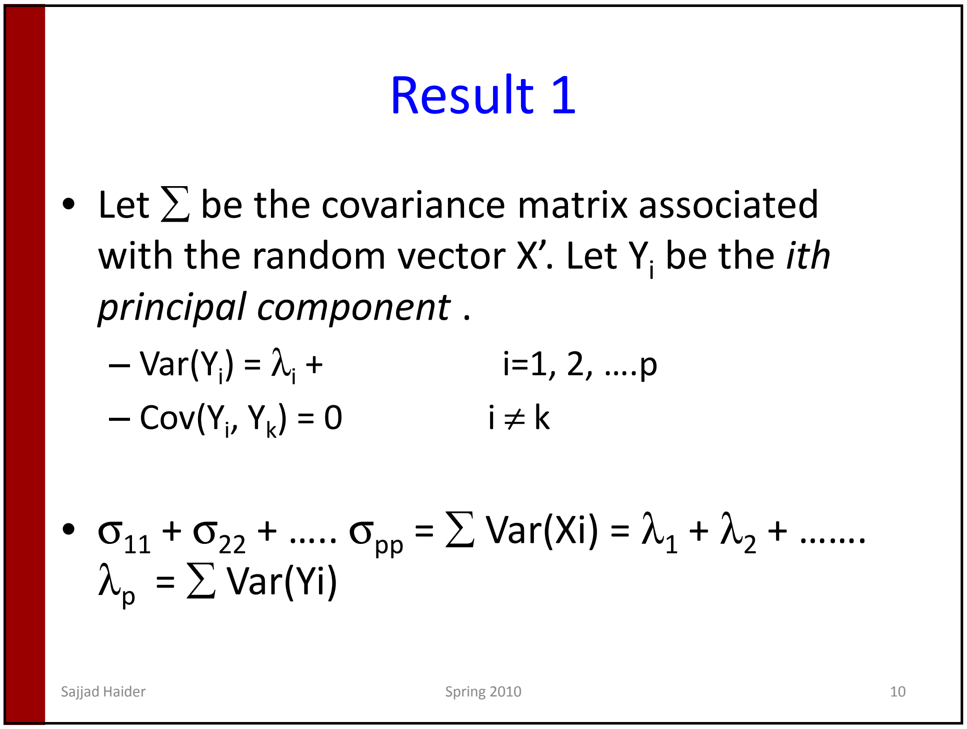
Practical Issues

- Because FA and PCA are exquisitely sensitive to the sizes of correlations, it is critical that honest, reliable correlations be employed.
- Sensitivity to outlying cases, problems created by missing data, and degradation of correlations between poorly distributed variables all plague FA and PCA.



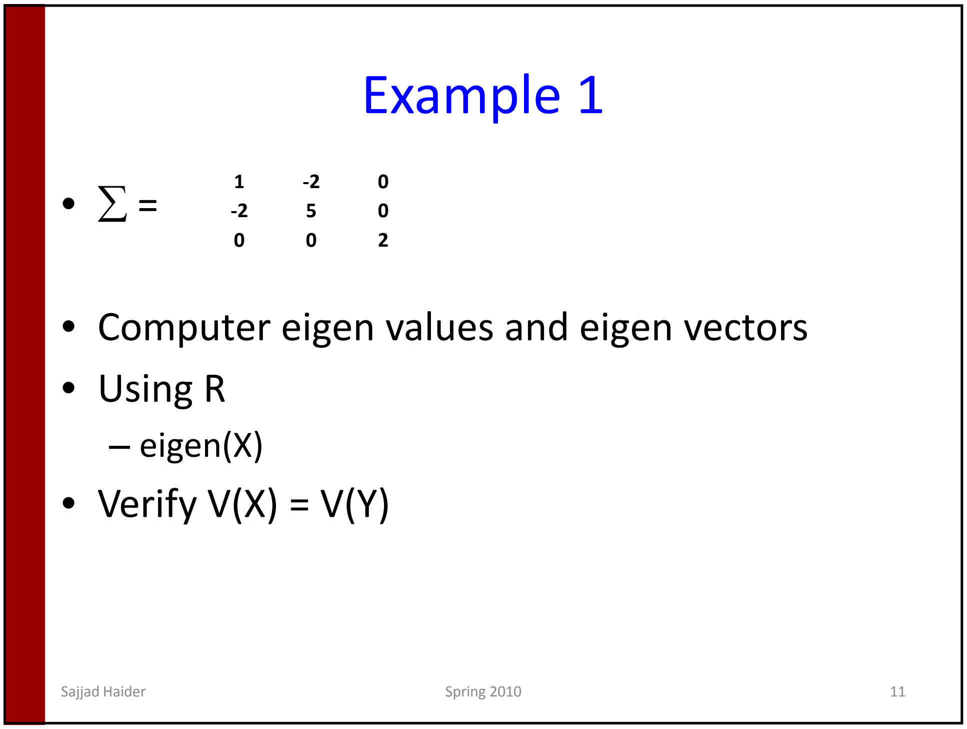
Normality

- As long as PCA and FA are used descriptively as convenient ways to summarize the relationships in a large set of observed variables, assumptions regarding the distributions of variables are not in force.
- If variables are normally distributed, the solution is enhanced. To the extent that normality fails, the solution is degraded but may still be worthwhile.
- However, multivariate normality is assumed when statistical inference is used to determine the number of factors. Multivariate normality is the assumption that all variables, and all linear combinations of variables, are normally distributed.



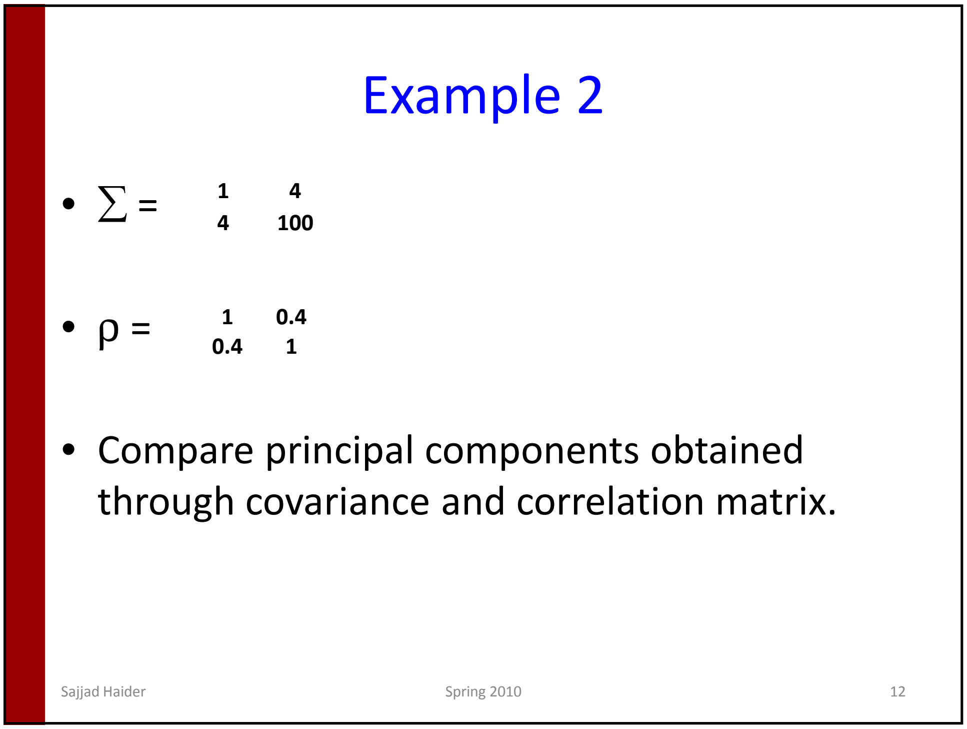
Result 1

- Let $\sum$ be the covariance matrix associated with the random vector X'. Let Y_i be the *ith principal component* .
  - $\text{Var}(Y_i) = \lambda_i +$ $i=1, 2, \ldots.p$
  - $\text{Cov}(Y_i, Y_k) = 0$ $i \neq k$
- $\sigma_{11} + \sigma_{22} + \ldots.. \sigma_{pp} = \sum \text{Var}(Xi) = \lambda_1 + \lambda_2 + \ldots\ldots. \lambda_p = \sum \text{Var}(Yi)$



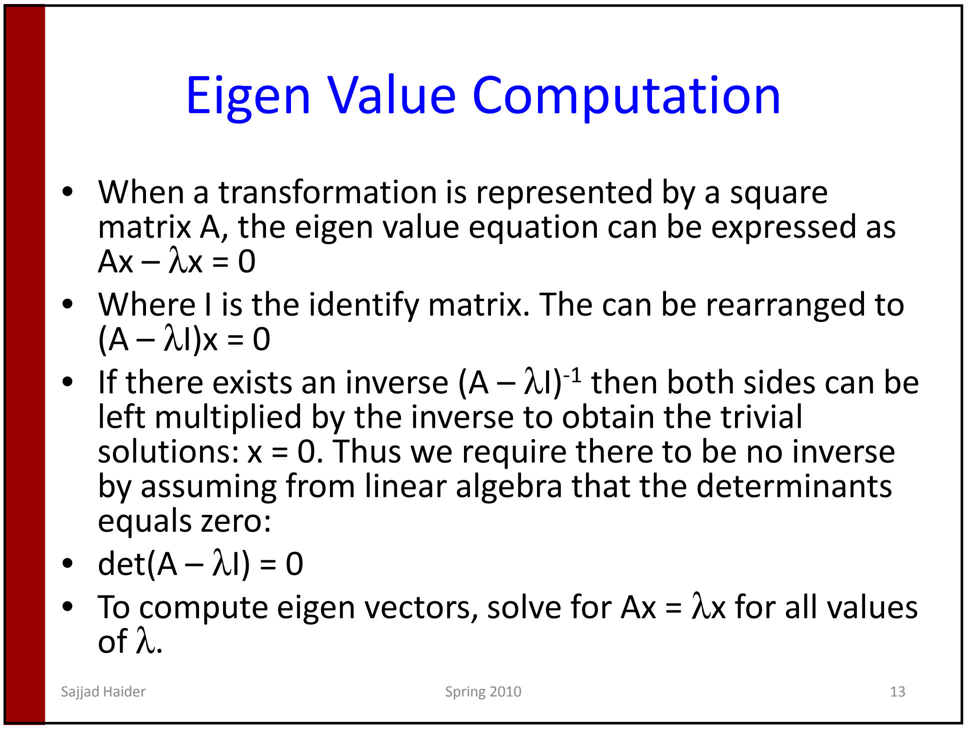
Example 1

- $\sum = \begin{matrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{matrix}$
- Computer eigen values and eigen vectors
- Using R
  - eigen(X)
- Verify V(X) = V(Y)



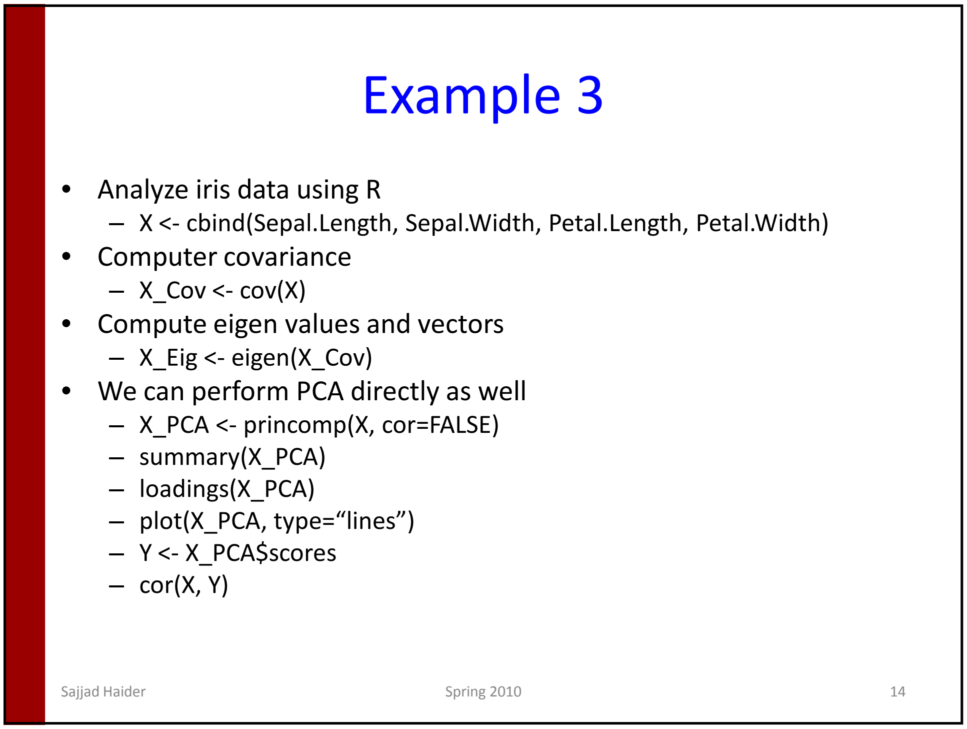
Example 2

- $\sum = \begin{matrix} 1 & 4 \\ 4 & 100 \end{matrix}$
- $\rho = \begin{matrix} 1 & 0.4 \\ 0.4 & 1 \end{matrix}$
- Compare principal components obtained through covariance and correlation matrix.



Eigen Value Computation

- When a transformation is represented by a square matrix A, the eigen value equation can be expressed as $Ax - \lambda x = 0$
- Where I is the identify matrix. The can be rearranged to $(A - \lambda I)x = 0$
- If there exists an inverse $(A - \lambda I)^{-1}$ then both sides can be left multiplied by the inverse to obtain the trivial solutions: $x = 0$. Thus we require there to be no inverse by assuming from linear algebra that the determinants equals zero:
- $\det(A - \lambda I) = 0$
- To compute eigen vectors, solve for $Ax = \lambda x$ for all values of λ.



Example 3

- Analyze iris data using R
  - X <- cbind(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width)
- Computer covariance
  - X_Cov <- cov(X)
- Compute eigen values and vectors
  - X_Eig <- eigen(X_Cov)
- We can perform PCA directly as well
  - X_PCA <- princomp(X, cor=FALSE)
  - summary(X_PCA)
  - loadings(X_PCA)
  - plot(X_PCA, type="lines")
  - Y <- X_PCA$scores
  - cor(X, Y)