Multivariate Strategies

See pp 3-4

Data Reduction or Simplification
Sorting and Grouping
Investigation of the dependence among variables
Prediction
Hypothesis Testing

Exercise for Sept 9 class: In a context of interest to you, suggest a hypothetical application of each of these strategies.

Some Basics

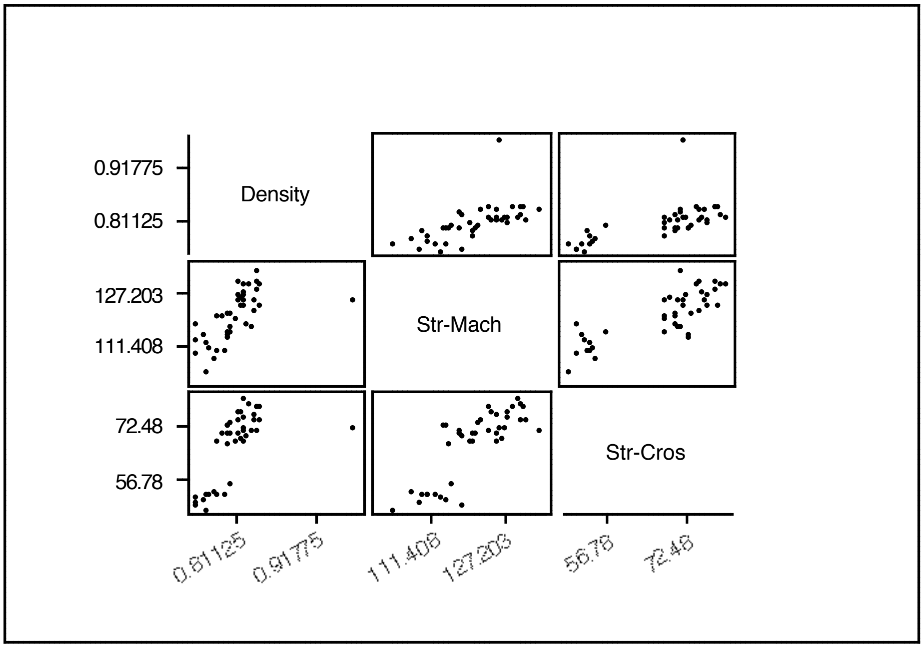
Definition of Sample Covariance and Sample Correlation. P 8. note n.

Means and covariances as summary stats of multivariate data set – when is it a good summary?

Outlier effect on correlation p 12. What would D&B have to be to make the correlation =0?

Ans (70,10) → (70,30)

Paper Quality Data – p 15 – Matrix Plot



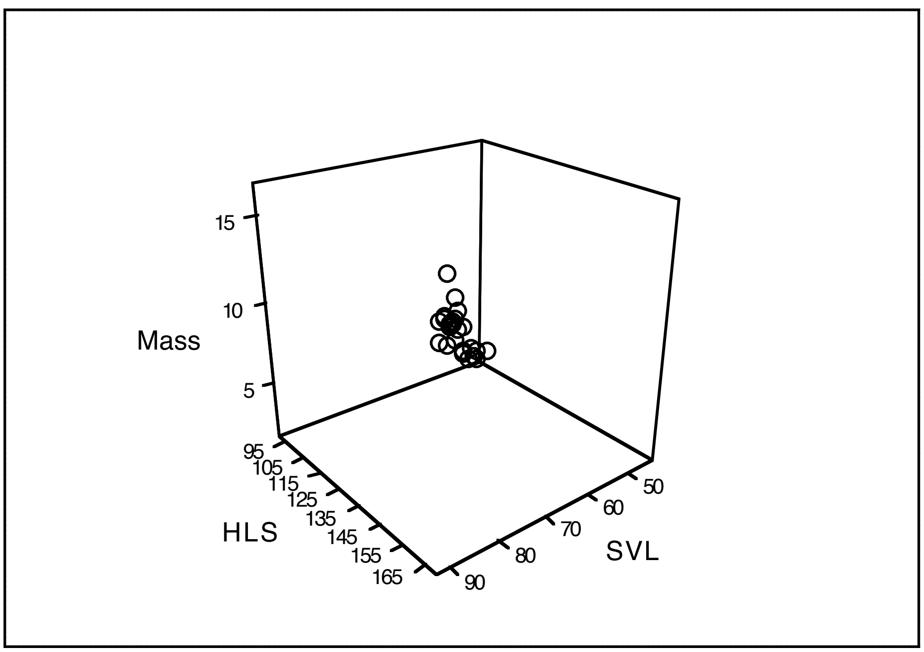
Any unusual patterns?   Outliers? Clusters?

Useful observation? Why?

Lower dimensional structure

Lizard data p 17

Wrong Perspectives:



Correlations (Pearson)

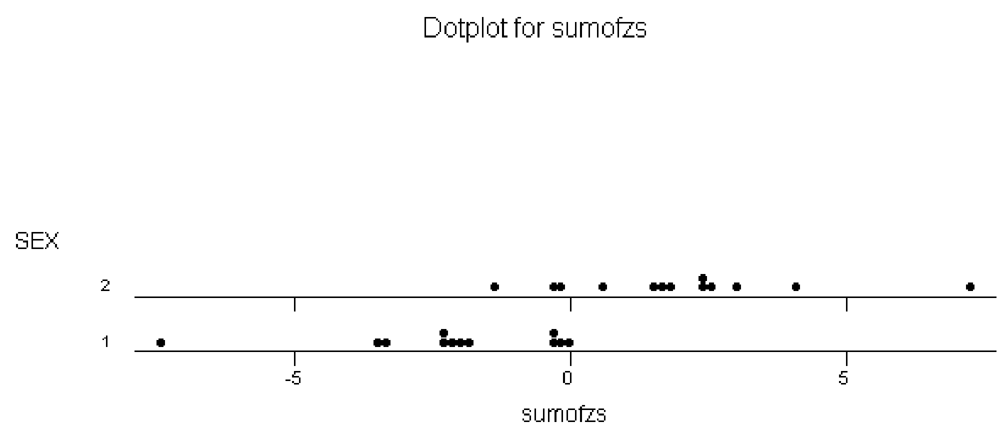
	Mass	SVL
SVL	0.967	
HLS	0.918	0.938

Form Index: sum of standardized values:

Correlations (Pearson)

	sumofzs	Mass	SVL
Mass	0.981		
SVL	0.988	0.967	
HLS	0.971	0.918	0.938

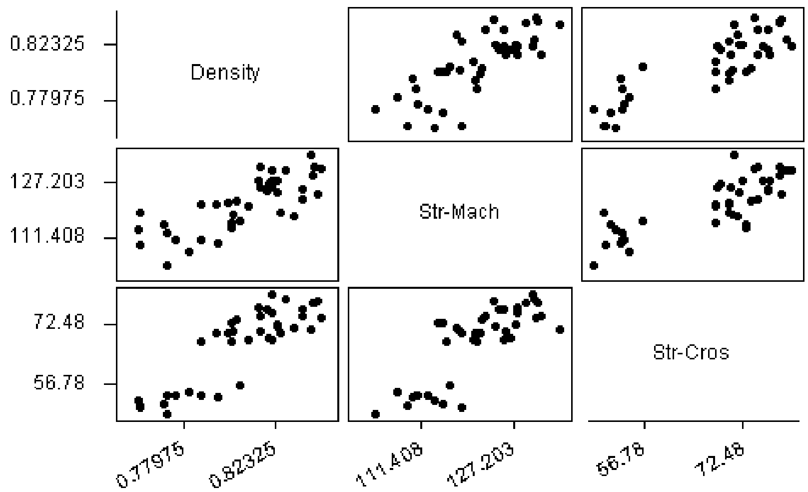
Explore Sex: Example of looking for structure in three dimensions



Brushing the paper-strength data – pp 20-21

Get rid of the outlier:

Result is



Paper Quality data – using rotation to visualize 3-D.

Other **ways to plot multivariate data** – see pp 24-30

Distance – Euclidean and Statistical – key concept.

Recall ordinary (Euclidean) distance formula: root sum square coord deviations

See Fig 1.20 : Consider distance from centroid.

If variables uncorrelated, just use Euclidean distance on standardized variables.

If correlated, transform to independence (rotate axes) and use above.

Distance between any two points is computed similarly (use diffs in cords).

Fig 1.23, Eqn 1-17 and 1-18 show how statistical distance in the uncorrelated situation can be generalized to the case of correlated variables. The only question is, how do we find the appropriate a11, a12, and a22 from data. Intutitively, the diagram suggests that the covariance matrix must be key, since it really determines the shape and extent of the scatter.   In Ch 2 we will see that the calculation of statistical distance depends entirely on the eigenvectors and eigenvalues of the covariance matrix. The eigenvectors determine the rotation of axes to achieve uncorrelated variables, and the eigenvalues give the variances in the directions of the eigenvectors.