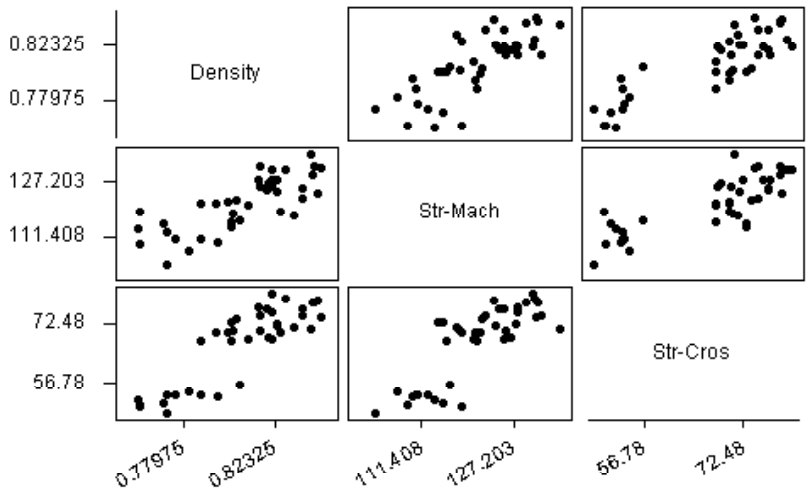
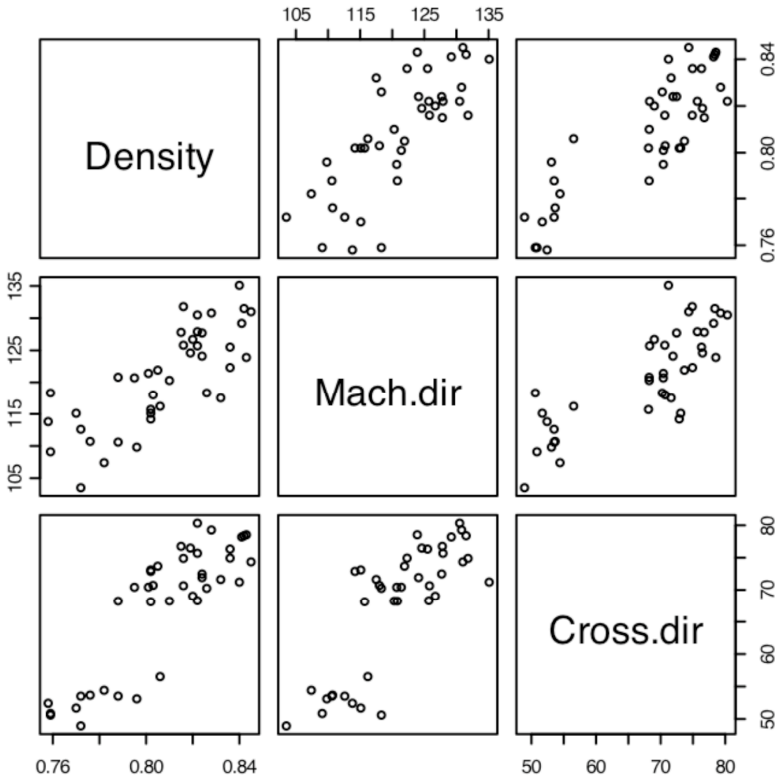
Paper-Quality Data (p 15 and matrix plot p 16)

Get rid of the outlier:

Result (Minitab Version):



R-version:



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

Other **ways to plot multivariate data** – see pp 24-30 Use of Splus or R.

Try

>stars(T1.2.df,lwd=1,key.loc=c(3,17.8)) to see the unusual points in the Paper Quality Data.

Distance – Euclidean and Statistical – key concept. (pp 30-37)

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

See Fig 1.20 (p 31) : Consider distance from centroid. Want to standardize variables?

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 (p 35) 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. Intuitively, 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.

One can project the data onto the eigenvectors to compute the new coordinates for the data.

Back up top review matrix algebra:

Some details (Ch 2)

Vector Representation and Arithmetic.

L_x = Length of a vector $x = (x'x)^{1/2}$ (Note x'x is a scalar).
Angle between a vector $x = (x_1,x_2)$ and the x_1 axis is $\cos^{-1}(x_1/L_x)$ (By definition of cos)
Inner Product of vectors x and y is denoted x'y and is scalar.
If θ is the angle between vectors x and y, then $\cos(\theta) = x'y/(L_xL_y)$
Generalizes to k dimensions. (not k variables!)

See example p 54

Recall definition of Linear Dependence
Vectors $x_1, x_2, \ldots, x_p$ are linearly dependent if you can express one as a linear function of the others. (See (2.7) p 54 bottom)

Projection of vector x on y – must be a multiple of the vector y – but what multiple?
Ans: x'y/y'y (see equation 2-8) so the projection is (x'y/y'y)y
If we want to write the projection of x on y as a multiple of a unit vector, we re-scale y to be unit by y/L_y and multiply it by $|x'y|/L_y$
Can infer that length of this projection is $|x'y|/y'y = L_x \cos(\theta)$

Matrices (p 55 ff)
Arithmetic: Sums, Products, Inverses and the Identity Matrix
Symmetric Matrices are Important for Statistics (Why?)
Orthogonal (Square) Matrices: $A' = A^{-1}$ so $A'A = I$
(Note: Orthonormal might be better, with Orthogonal meaning A'A=Diagonal – but we will stick to the terminology of the book).

Eigenvalues and Eigenvectors

(Synonyms Latent and Characteristic have other meanings as well – confusing)

Square Matrix A – x is Eigen vector of A if $Ax=\lambda x$ where λ is a scalar. λ is called an eigenvalue of A corresponding to the eigenvector x. There are usually many such x and λ, but never more than the rank of A ($\leq$ the dimension of A).

Now concentrate on (square) symmetric matrices (like covariance matrices) – see theorem p 61. Eigenvectors are mutually perpendicular unless some eigenvalues are multiple.
By convention, usually specify eigenvectors to have unit length (wlog see definition).

Positive Definite Matrices

A symmetric matrix A is Positive Definite if $x'Ax > 0$ for all $x \neq 0$
Can show this is equivalent to all eigenvalues >0 (p 64)

Statistical Distance of vector x from vector 0 (origin) has the form x'Ax where A is pos definite.
From x to μ is $(x-\mu)'A(x-\mu)$. But we still have to make A specific to call it statistical distance.

To find appropriate A: See p 66 Fig 2.6 If the ellipse shown were a contour of the bivariate normal density, then we would find e_1, e_2, λ_1 and λ_2 from the eigenanalysis of the Σ^{-1} where Σ is the covariance matrix. So the distance of points from the origin would be $x'\Sigma^{-1}x$ (In the diagram, the variables are centered so the mean is the origin). This generalizes to p dimensions (See Theorem bottom p 81) . It implies that if you keep extracting maximal variance projections, each one perpendicular to the previous one, then the eigenvalues will be the maximal variances at each stage, and the eigenvectors will be the directions that the projections must be taken.

The interpretation depends on the multivariate normality, but the procedure does not. The means and covariances of the data will produce statistical distances from the centroid for every "case".

Σ can be expressed as $E[(x-\mu)'(x-\mu)]$ and is estimated by the equivalent sample formula (p 125). Because Σ is positive definite usually, there is a spectral representation based on the eigenanalysis (see p 67 , (2-21)). This represeantation shows that the statistical distance is indeed found by an orthogonal rotation of the axes followed by the Euclidean distance in the new coordinates. The A in this formula can be replaced by Σ^{-1} or its sample estimate.

Exercise: Choose any multivariate data set in the text with quantitative variables. Compute the dotplot of statistical distances for the cases in the data set. Hand in your result on Wed, Sept 14 at class.