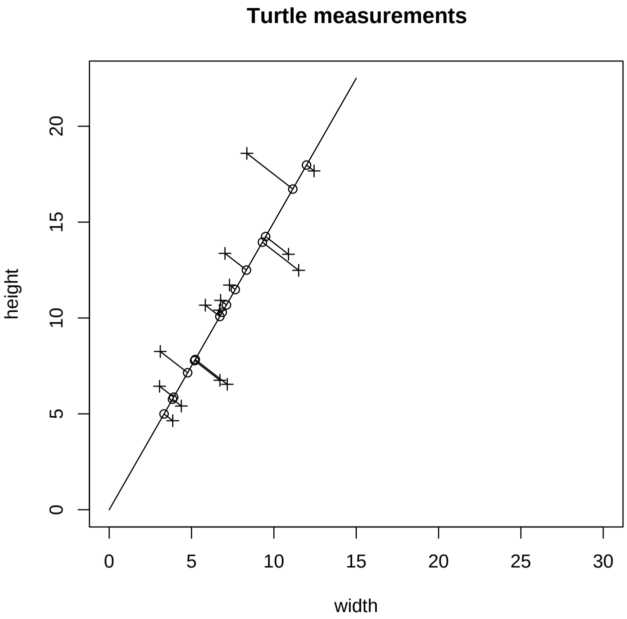
3/2/2010 36-402/608 ADA-II H. Seltman
Handout #14: Principal Components Analysis

1. Overview

   (a) PCA (principal component analysis) is a dimension reduction technique, usable for multiple correlated explanatory or response variables, that results in a smaller set of uncorrelated "principal component" variables that capture most of the information in the original variables.

   (b) The new variables are often meaningful and simpler to understand and work with than the original variables

2. Example of the main idea in two dimensions: In a study of turtle diets (for one species), two outcomes are measured: the shell length and width.



   The line drawn is the "first PC" and the corresponding spacing of the circles has the maximum variance of all potential lines.

3. Conceptual algorithm: Find the direction in p-space that maximizes the variance of the projection onto that new, rotated axis. Repeatedly, find the next direction that maximizes the variance of the projection onto that axis among all directions perpendicular to the preceding axes. The coordinates of the data points on the new axes are the 1st, 2nd, etc. principal components of the data.

4. Practical algorithm: Eigenanalysis (or SVD) of the covariance matrix of the data. Variances of the principal components are the eigenvalues. Directions are the eigenvectors.

5. Other details

   (a) The direction of each principal component is expressed as its **loadings**, which are the coefficients that defines a linear combination of each original variable that produces the value of the given principal component (for each subject).

   (b) Although there are an infinite set of loadings that define the same direction, a unique set is defined using the **constraint** that the sum of the squares of the loadings for each p.c. is 1. This also assures that the sum of the variances of all p.c.s equals the sum of the variances of the original variables.

   (c) Whether a given p.c.s loading is chosen vs. **-1** times that (entire) loading is arbitrary and meaningless.

   (d) A common eyeball test of how many p.c.s to chose as a data reduction is the **scree plot** of the fraction of the total variance for each p.c. against p.c. number. Look for a "breakpoint".

   (e) It may be reasonable to **use a similar, more meaningful set of coefficients** for further analysis, e.g., instead of (0.54, 0.62, 0.57), use the mean.

   (f) Since the **p.c.s are uncorrelated**, if the reduction is for a set of response variables, separate analyses are reasonable.

   (g) Since the p.c.s are uncorrelated, if the reduction is for a set of explanatory variables, removal of one variable will not affect the coefficients of the other variables.

   (h) PCA results change if the **units** of some (but not all) variables are changed, e.g., cm. to mm.

   (i) For disparate measures, PCA is usually done on **standardized variables** (Z-score). Note that this has a large degree of arbitrariness if some variables are artificially restricted, e.g., only allow ages 20-40.

6. Related topics

   (a) Rather than looking for inherently meaningful directions in the measured variables, "factor analysis" focuses on inferring unmeasured (latent) variables which then determine the measured variables. E.g., performance on a the math SAT may be determined by a general math ability quality, an ability to relate word problem text to equations, algebra skills, and geometry skills. Then each problem requires some weightings of each of these skills. A key quantity is "communality": variance of the observed variables accounted for by a given factor.

7. Breakout and Discussion