Use of Covariance in Distance

- Similarities between cup

- Suppose we measure cup-height 100 times and diameter only once
  - height will dominate although 99 of the height measurements are not contributing anything
- They are very highly correlated
- To eliminate redundancy we need a *data-driven method*
  - approach is to not only to standardize data in each direction but also to use covariance between variables

Covariance between two Scalar Variables

Sample means

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}\left(x(i) - \bar{x}\right)\left(y(i) - \bar{y}\right)$$

- A scalar value to measure how x and y vary together
- Obtained by
  - multiplying for each sample its mean-centered value of x with mean-centered value of y
  - and then adding over all samples
- Large positive value
  - if large values of x tend to be associated with large values of y and small values of x with small values of y
- Large negative value
  - if large values of x tend to be associated with small values of y
- With d variables can construct a d x d matrix of covariances
  - Such a covariance matrix is symmetric.

For Vectors: Covariance Matrix and Data Matrix

- Let X = n **x** d data matrix
- Rows of X are the data vectors $x(i)$
- Definition of covariance:

$$Cov(i,j) = \frac{1}{n}\sum_{k=1}^{n}\left(x_k(i) - \bar{x}\right)\left(y_k(i) - \bar{y}\right)$$

- If values of X are mean-centered
  - i.e., value of each variable is relative to the sample mean of that variable
  - then $V = X^T X$ is the d x d covariance matrix

Correlation Coefficient

Value of Covariance is dependent upon ranges of x and y

Dependency is removed by

dividing values of x by their standard deviation

and values of y by their standard deviation

$$\rho(X,Y) = \frac{\sum_{i=1}^{n}(x(i)-\bar{x})(y(i)-\bar{y})}{\sigma_x \sigma_y}$$

With p variables, can form a d x d correlation matrix

Correlation Matrix

Housing related variables across city suburbs ($d=11$)

11 x 11 pixel image (White 1, Black -1)
Columns 12-14 have values -1,0,1 for pixel intensity reference
Remaining represent corrrelation matrix

Variables 3 and 4 are highly negatively correlated with Variable 2
Variable 5 is positively correlated with Variable 11
Variables 8 and 9 are highly correlated

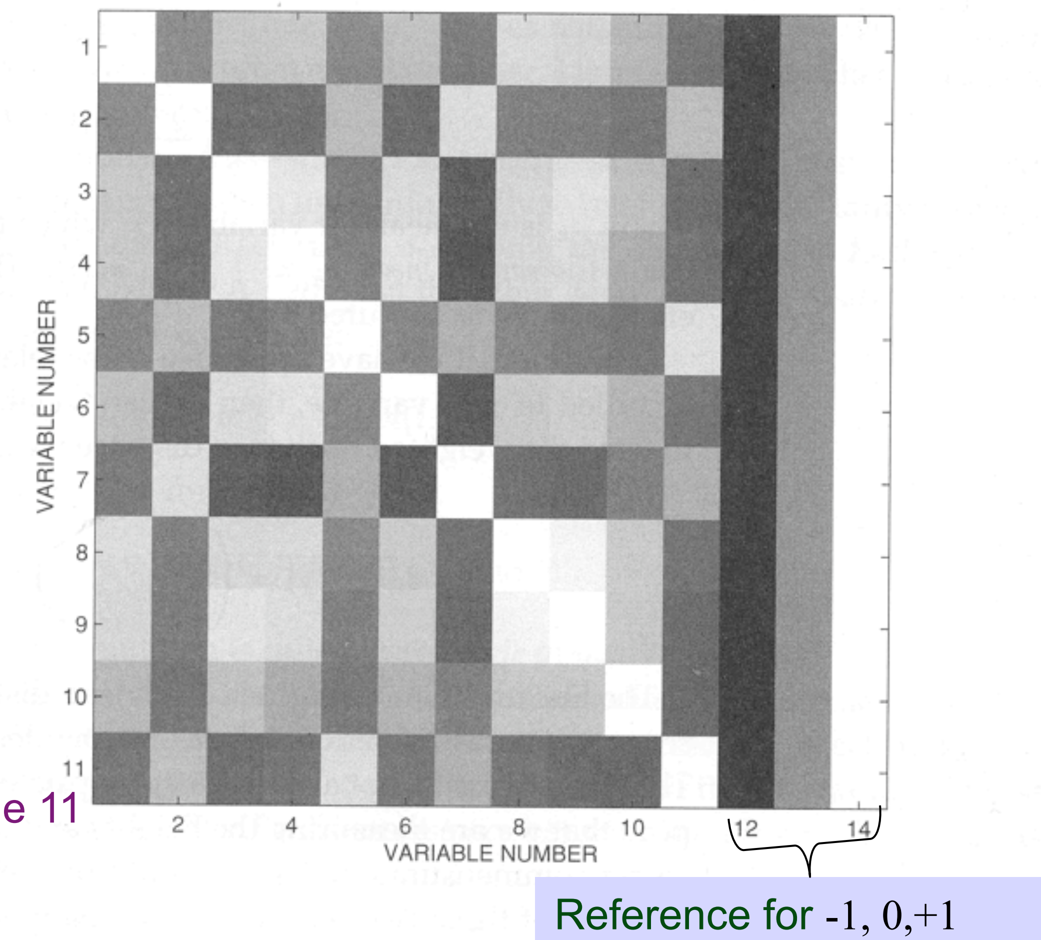

Figure 2.1 A sample correlation matrix plotted as a pixel image. White corresponds to +1 and black to -1. The three rightmost columns contain values of -1, 0, and +1 (respectively) to provide a reference for pixel intensities. The remaining 11×11 pixels represent the 11×11 correlation matrix. The data come from a well-known data set in the regression research literature, in which each data vector is a suburb of Boston and each variable represents a certain general characteristic of a suburb. The variable names are (1) per-capita crime rate, (2) proportion of area zoned for large residential lots, (3) proportion of non-retail business acres, (4) nitric oxide concentration, (5) average number of rooms perdwelling, (6) proportion of pre-1940 homes, (7) distance to retail centers index, (8) accessibility to highways index, (9) property tax rate, (10) pupil-to-teacher ratio, and (11) median value of owner-occupied homes.

Incorporating Covariance Matrix in Distance

Mahanalobis Distance between samples $x(i)$ and $x(j)$ is:

$$d_M(x(i), x(j)) = \Big[\underbrace{\big(x(i) - x(j)\big)^T}_{1 \times d} \underbrace{\sum\nolimits^{-1}}_{d \times d} \underbrace{(x(i) - x(j))}_{d \times 1}\Big]^{\frac{1}{2}}$$

T is transpose
Σ is d x d covariance matrix
Σ^{-1} standardizes data relative to Σ

Matrix multiplication yields a scalar value

d_M discounts the effect of several highly correlated variables