Principal Components

- Often, a multivariate response has relatively few dominant *modes of variation*.

- E.g. length, width, and height of turtle shells: overall size is a dominant mode of variation, making all three responses larger or smaller simultaneously.

- Principal components analysis (PCA) explores modes of variation suggested by a variance-covariance matrix.
• Identifying the modes of variation can lead to:
  
  – new insights and interpretation;

  – reduction in dimensionality and collinearity.

• Often used to prepare data for further analysis, e.g. in regression analysis (Principal Components Regression).
• Basic idea: find linear combinations of responses that have most of the variance.

• More precisely: given \( X \) with \( E(X) = 0 \) and \( \text{Cov}(X) = \Sigma \), find \( a_1 \) to maximize \( \text{Var}(a_1'X) \), subject to \( a_1'a_1 = 1 \).

• Notes:
  
  – \( a_1 \) must be constrained, otherwise the variance could be made arbitrarily large just by making \( a_1 \) large.
  
  – Why this constraint? Because the problem has a convenient solution.
Solution:

\[ \text{Var}(a'_1 X) = a'_1 \Sigma a_1, \]

which is maximized at \( a_1 = e_1 \), the first eigenvector of \( \Sigma \);

The maximized value is \( \lambda_1 \), the associated eigenvalue.

Often, the elements of \( e_1 \) are all positive and similar in magnitude \( \Rightarrow \) a mode in which all responses vary together:

- not often interesting;

- a useful summary or composite variable.
• What about other modes of variation? Johnson & Wichern: find $a_2$ to maximize $\text{Var}(a'_2X)$, subject to:

- $a'_2a_2 = 1$;
- $\text{Cov}(a'_1X, a'_2X) = 0$.

• Solution:

- $a_2 = e_2$, the second eigenvector of $\Sigma$;
- The maximized value is $\lambda_2$, the associated eigenvalue.

• Similarly modes 3, 4, . . . , $p$. 
Alternative Approach: Data Compression

- Observer sees $X$ but communicates only $Y = a'X$.

- Receiver uses $X^* = bY$ to approximate $X$.

- Error is $X - X^* = (I - ba')X$.

- Error covariance matrix is $\text{Cov}(X - X^*) = (I - ba')\Sigma(I - ab')$.

- *Total* error variance is
  \[
  \text{trace}[\text{Cov}(X - X^*)] = \text{trace}(\Sigma) - a'\Sigma b - b'\Sigma a + b'ba'\Sigma a.
  \]
• For a given $b$, total error variance is minimized by 

$$a = \frac{b}{b'bb'}.$$ 

• Minimum for a given $b$ is 

$$\text{trace}(\Sigma) - b'\Sigma b \frac{b'}{b'b}.$$ 

• Best $b$ is $e_1$, and then best $a = e_1$ also. 

• Best (i.e., minimum) total error variance is 

$$\text{trace}(\Sigma) - \lambda_1 = \lambda_2 + \lambda_3 + \cdots + \lambda_p.$$
Higher dimensions: suppose observer can communicate \( Y = (Y_1, Y_2)' = A'X \), and receiver uses \( X^* = BY \) to approximate \( X \).

Similar math: best \( A = B = (e_1, e_2) \), with total error

\[
\text{trace}(\Sigma) - \lambda_1 - \lambda_2 = \lambda_3 + \lambda_4 + \cdots + \lambda_p.
\]

Similarly for \( k \) channels, \( k < p \): proportion of total variance “explained” (i.e., communicated) is

\[
\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_k + \lambda_{k+1} + \cdots + \lambda_p}.
\]