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 $\mathbf{X}$ with $\mathbb{E}(\mathbf{X}) = \mathbf{0}$ and $\text{Cov}(\mathbf{X}) = \mathbf{\Sigma}$, find $\mathbf{a}_1$ to maximize $\text{Var}(\mathbf{a}_1^\prime \mathbf{X})$, subject to $\mathbf{a}_1^\prime \mathbf{a}_1 = 1$.

Notes:
- $\mathbf{a}_1$ must be constrained, otherwise the variance could be made arbitrarily large just by making $\mathbf{a}_1$ large.
- Why this constraint? Because the problem has a convenient solution.
Solution:

\[ \text{Var} \left( a_1' X \right) = a_1' \Sigma a_1, \]

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 \textit{composite} variable.
What about other modes of variation? Find $a_2$ to maximize $\text{Var}(a'_2 X)$, subject to:

- $a'_2 a_2 = 1$;
- $\text{Cov}(a'_1 X, a'_2 X) = 0$.

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 \( \mathbf{X} \) but communicates only \( Y = a'\mathbf{X} \).
- Receiver uses \( \mathbf{X}^* = \mathbf{b}Y \) to approximate \( \mathbf{X} \).
- Error is \( \mathbf{X} - \mathbf{X}^* = (\mathbf{I} - \mathbf{ba}')\mathbf{X} \).
- Error covariance matrix is \( \text{Cov}(\mathbf{X} - \mathbf{X}^*) = (\mathbf{I} - \mathbf{ba}')\Sigma(\mathbf{I} - \mathbf{ab}') \).
- Total error variance is

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

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

Minimum for a given $b$ is

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

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

- Recall: PCA is the solution to a data compression problem, where “error” is quantified by total error variance.
- Question: is “total variance” appropriate?
- Variables in different units must be scaled.
- Variables in the same units but with very different variances are usually scaled.
Simplest scaling: divide each variable by its standard deviation ⇒ covariances are *correlations*.

In other words: use eigen structure of *correlation* matrix $R$, not *covariance* matrix $\Sigma$. 
PCA for some Special Cases

- Diagonal matrix: if

\[ \Sigma = \begin{bmatrix}
\sigma_{1,1} & 0 & \cdots & 0 \\
0 & \sigma_{2,2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sigma_{p,p}
\end{bmatrix} \]

then the principal components are just the original variables.
Compound symmetry: if

\[ \Sigma = \begin{bmatrix} \sigma^2 & \rho \sigma^2 & \ldots & \rho \sigma^2 \\ \rho \sigma^2 & \sigma^2 & \ldots & \rho \sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho \sigma^2 & \rho \sigma^2 & \ldots & \sigma^2 \end{bmatrix} \]

then (if \( \rho > 0 \)):

- \( \lambda_1 = 1 + (p - 1) \rho \) and \( \mathbf{e}_1 = p^{-1/2} (1, 1, \ldots, 1)' \);
- \( \lambda_k = 1 - \rho, \ k > 1; \)
- \( \mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_p \) are an arbitrary basis for the rest of \( \mathbb{R}^p \).
- If \( \rho < 0 \) the order is reversed, but note that \( \rho \) must satisfy \( 1 + (p - 1) \rho \geq 0 \Rightarrow \rho \geq -1/(p - 1) \).
Principal Components

- Time series ($1^{\text{st}}$ order autoregression):

\[
\Sigma = \begin{bmatrix}
\sigma^2 & \phi \sigma^2 & \ldots & \phi^{p-1} \sigma^2 \\
\phi \sigma^2 & \sigma^2 & \ldots & \phi^{p-2} \sigma^2 \\
\vdots & \vdots & \ddots & \vdots \\
\phi^{p-1} \sigma^2 & \phi^{p-2} \sigma^2 & \ldots & \sigma^2
\end{bmatrix}
\]

- No closed form, but for large $p$ the eigen vectors are like sines and cosines.
Sample PCA

- Essentially the eigen analysis of $S$ (or $R$):

\[ S\hat{e}_k = \hat{\lambda}_k \hat{e}_k, \]

and

\[ \hat{y}_k = X_{\text{dev}} \hat{e}_k, \]

where

\[ X_{\text{dev}} = X - \frac{1}{n}11'X = \left(I - \frac{1}{n}11'\right)X \]
Note:

\[ S = \frac{1}{n-1} \mathbf{X}_{\text{dev}}' \mathbf{X}_{\text{dev}} = \left( \frac{1}{\sqrt{n-1}} \mathbf{X}_{\text{dev}} \right)' \left( \frac{1}{\sqrt{n-1}} \mathbf{X}_{\text{dev}} \right) \]

**Singular value decomposition:**

\[ \frac{1}{\sqrt{n-1}} \mathbf{X}_{\text{dev}} = \mathbf{U} \mathbf{D} \mathbf{V}' \]

where \( \mathbf{U} \) and \( \mathbf{V} \) have orthonormal columns and \( \mathbf{D} \) is diagonal (but may not be square).

The diagonal entries of \( \mathbf{D} \) are the square roots of the largest \( p \) eigenvalues of both \( (n-1)^{-1} \mathbf{X}_{\text{dev}}' \mathbf{X}_{\text{dev}} = S \) and \( (n-1)^{-1} \mathbf{X}_{\text{dev}} \mathbf{X}_{\text{dev}}' \).

The columns of \( \mathbf{V} \) are the eigenvectors of \( \mathbf{X}_{\text{dev}}' \mathbf{X}_{\text{dev}} \).
Also

\[ \mathbf{X}_{\text{dev}} \mathbf{V} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_p] = \left( \sqrt{n - 1} \right) \mathbf{UD} \]

so the singular value decomposition of \((n - 1)^{-1/2} \mathbf{X}_{\text{dev}}\) provides all the details of the sample principal components:

- the coefficients \(\mathbf{V}\);
- the values \(\mathbf{UD}\).

Similarly, if \(\mathbf{X}^*\) is \(\mathbf{X}_{\text{dev}}\) with its columns normalized (sum of squares = 1), then

\[ \mathbf{R} = \mathbf{X}^*/\mathbf{X}^*, \]

and the singular value decomposition of \(\mathbf{X}^*\) gives the PCA of \(\mathbf{R}\).
Example: 5 stocks

- **DU PONT E I DE NEM (NYSE:DD)** (a former Dow Industrials stock)
- **HONEYWELL INTL INC (NYSE:HON)** (a former Dow Industrials stock)
- **EXXON MOBIL CP (NYSE:XOM)** (a Dow Industrials stock)
- **CHEVRON CORP (NYSE:CVX)** (a Dow Industrials stock)
- **DOW CHEMICAL (NYSE:DOW)** (former Dow stock)
SAS `proc princomp` program and output.

R code and graphs for an updated set of stocks: DD, HON, and XOM, plus MSFT (Microsoft) and WMT (Walmart).

```r
stocksPCAcor = prcomp(stocks(), scale. = TRUE);
print(stocksPCAcor);
plot(stocksPCAcor);
biplot(stocksPCAcor);

stocksPCAcov = prcomp(stocks());
print(stocksPCAcov);
plot(stocksPCAcov);
biplot(stocksPCAcov);
```
Biplot for standardized stock returns
stocksPCA cov

Variances

0 5 10 15

10

15
Principal Components

Biplot for unstandardized stock returns

NC STATE UNIVERSITY
How Many Components?

- How many components are important?
  - No definitive answer in the PCA framework.
  - The *factor analysis* model allows maximum likelihood estimation, hence hypothesis testing.
- For insight, use all that have a substantive interpretation.
- For other uses such as regression, we need an objective rule.
▶ Use “scree” plot (plot of eigenvalues), and look for an “elbow”; very subjective.
▶ Simple rule of thumb: use all eigenvalues larger than the average.
  ▶ Note: for the correlation matrix, the average is always 1, so the rule is: use all eigenvalues $> 1$.
▶ Overland and Preisendorfer (1982) proposed a rule ("Rule N") based on comparison of observed eigenvalues with the distribution of eigenvalues in the case of iid $N(0, \sigma^2)$.
Suppose that the rows of the data matrix $X$ are a random sample of size $n$ from $N_p(\mu, \Sigma)$.

Assume that $\Sigma$ has distinct eigenvalues

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0.$$  

Then, approximately for large $n$,

$$\sqrt{n} \left( \hat{\lambda} - \lambda \right) \sim N_p \left[ 0, 2 \times \text{diag} (\lambda^2) \right].$$
In other words, $\hat{\lambda}_i$ and $\hat{\lambda}_k$ are approximately independent for $k \neq i$, and

$$\hat{\lambda}_i \sim N\left(\lambda_i, \frac{2\lambda_i^2}{n}\right).$$

Note: if

$$\frac{n\hat{\lambda}}{\lambda} \sim \chi_n^2$$

then similarly, approximately,

$$\hat{\lambda} \sim N\left(\lambda, \frac{2\lambda^2}{n}\right).$$
So we could also state that, approximately,

\[
\frac{n\hat{\lambda}_i}{\lambda_i} \sim \chi_n^2.
\]

Simulations suggest that this is a better approximation for small \( n \), if the eigenvalues are well separated.

Asymptotics suggest that the degrees of freedom for \( \hat{\lambda}_i \) could be \( n - i + 1 \) instead of \( n \).

Also \( \sqrt{n}(\hat{e}_i - e_i) \) is approximately \( N_p (0, E_i) \), where

\[
E_i = \lambda_i \sum_{k=1}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} \times e_k e_k'.
\]
Test for equal correlations:

\[ H_0 : \mathbf{R} = \begin{bmatrix}
1 & \rho & \ldots & \rho \\
\rho & 1 & \ldots & \rho \\
\vdots & \vdots & \ddots & \vdots \\
\rho & \rho & \ldots & 1
\end{bmatrix} \]

against the alternative of a general (unstructured) \( \mathbf{\Sigma} \) or \( \mathbf{R} \).

- \( H_0 \) is equivalent to equality of all eigenvalues of \( \mathbf{R} \) but one.
- Likelihood ratio test (and Lawley’s test) give large-sample \( \chi^2 \) statistic with \((p + 1)(p - 2)/2\) d.f.
Other related tests of interest:

- *compound symmetry* is a variance-components structure; it is stronger, requiring equal correlations and equal variances;
- equivalent to equality of all eigenvalues of $\Sigma$ but one.
- *sphericity* is the necessary and sufficient condition for univariate repeated measures to be valid; it is a weaker condition: $\sigma_{i,j} = (\sigma_{i,i} + \sigma_{j,j})/2 - \tau^2$.
- no simple characterization in terms of eigen structure.