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 $\Rightarrow$ covariances are *correlations*.

- In other words: use eigen structure of *correlation* matrix $\mathbf{R}$, not *covariance* matrix $\mathbf{\Sigma}$.

PCA for some Special Cases

- Diagonal matrix: if

$$\Sigma = \begin{bmatrix} \sigma_{1,1} & 0 & \dots & 0 \\ 0 & \sigma_{2,2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{p,p} \end{bmatrix}$$

  then the principal components are just the original variables.

- Compound symmetry: if

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \dots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \dots & \rho\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \dots & \sigma^2 \end{bmatrix}$$

  then (if $\rho > 0$):

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

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

$$\Sigma = \begin{bmatrix} \sigma^2 & \phi\sigma^2 & \dots & \phi^{p-1}\sigma^2 \\ \phi\sigma^2 & \sigma^2 & \dots & \phi^{p-2}\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \phi^{p-1}\sigma^2 & \phi^{p-2}\sigma^2 & \dots & \sigma^2 \end{bmatrix}$$

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

Sample PCA

- Essentially the eigen analysis of $\mathbf{S}$ (or $\mathbf{R}$):

$$\mathbf{S}\widehat{\mathbf{e}}_k = \widehat{\lambda}_k \widehat{\mathbf{e}}_k,$$

and

$$\widehat{\mathbf{y}}_k = \mathbf{X}_{\mathsf{dev}} \widehat{\mathbf{e}}_k,$$

where

$$\mathbf{X}_{\mathsf{dev}} = \mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{X}$$

- Recall:

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

- *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 $(n-1)^{-1}\mathbf{X}'_{\text{dev}}\mathbf{X}_{\text{dev}} = \mathbf{S}$ 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

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

  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

$$\mathbf{R} = \mathbf{X}^{*\prime}\mathbf{X}^*,$$

  and the singular value decomposition of $\mathbf{X}^*$ gives the PCA of $\mathbf{R}$.

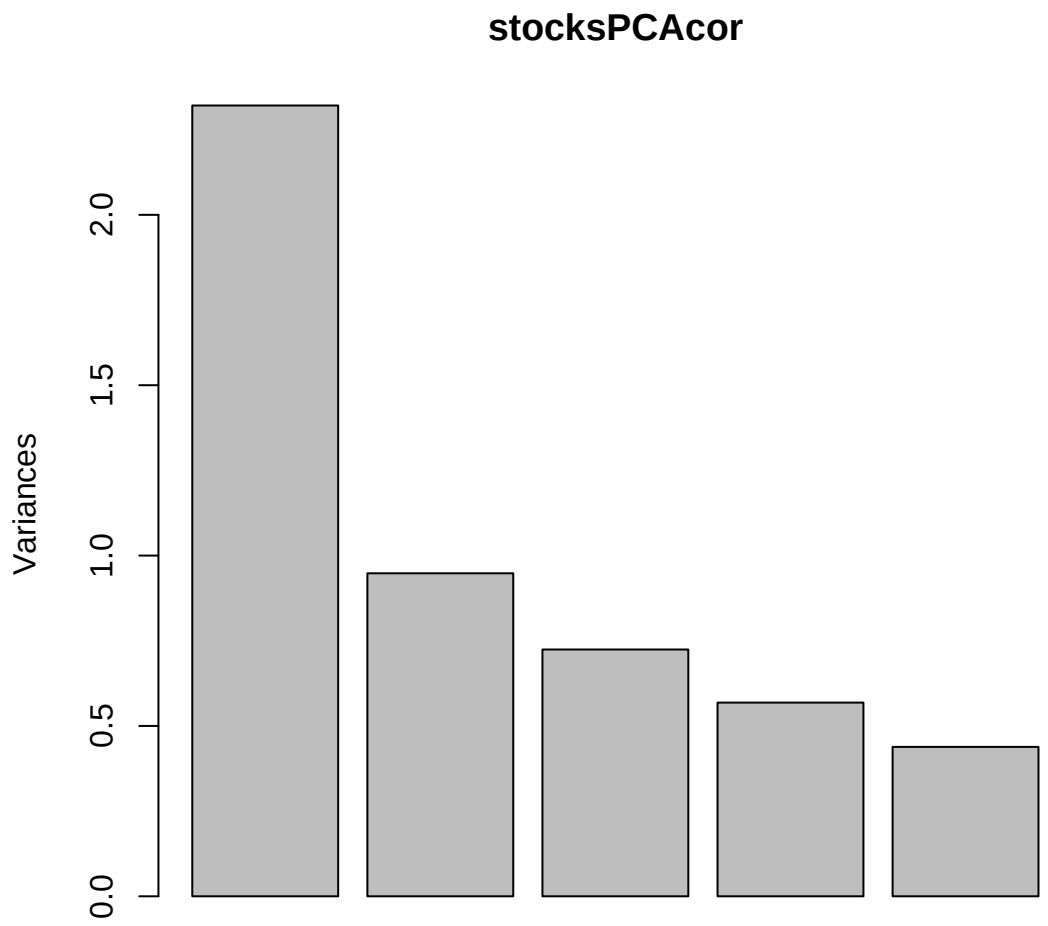
- Example: 5 stocks
  - DU PONT E I DE NEM (NYSE:DD) (a former Dow Industrials stock)
  - HONEYWELL INTL INC (NYSE:HON) (a former Dow Industrials stock)
  - EXXON MOBIL CP (NYSE:XOM) (a Dow Industrials stock)
  - CHEVRON CORP (NYSE:CVX) (a Dow Industrials stock)
  - DOW CHEMICAL (NYSE:DOW) (former Dow stock)

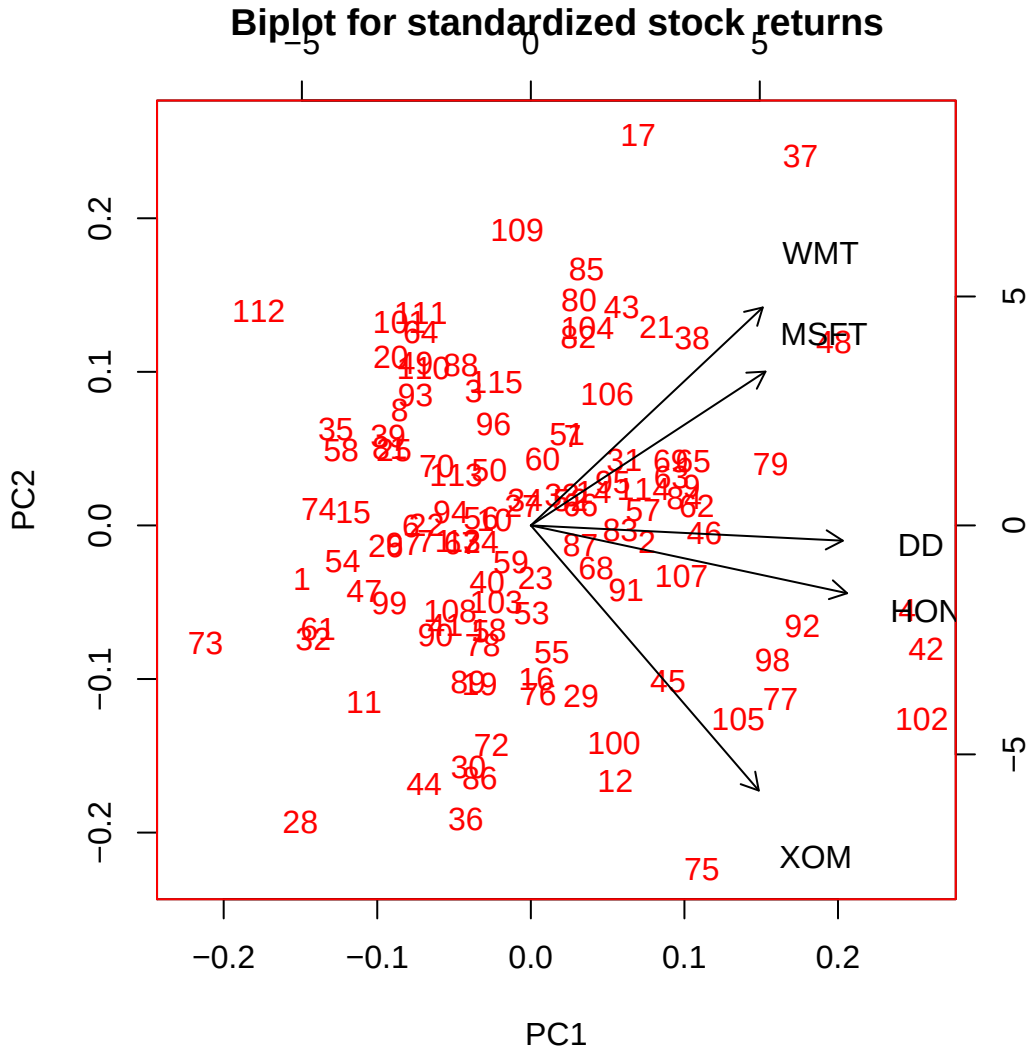
- SAS `proc princomp` program and output.

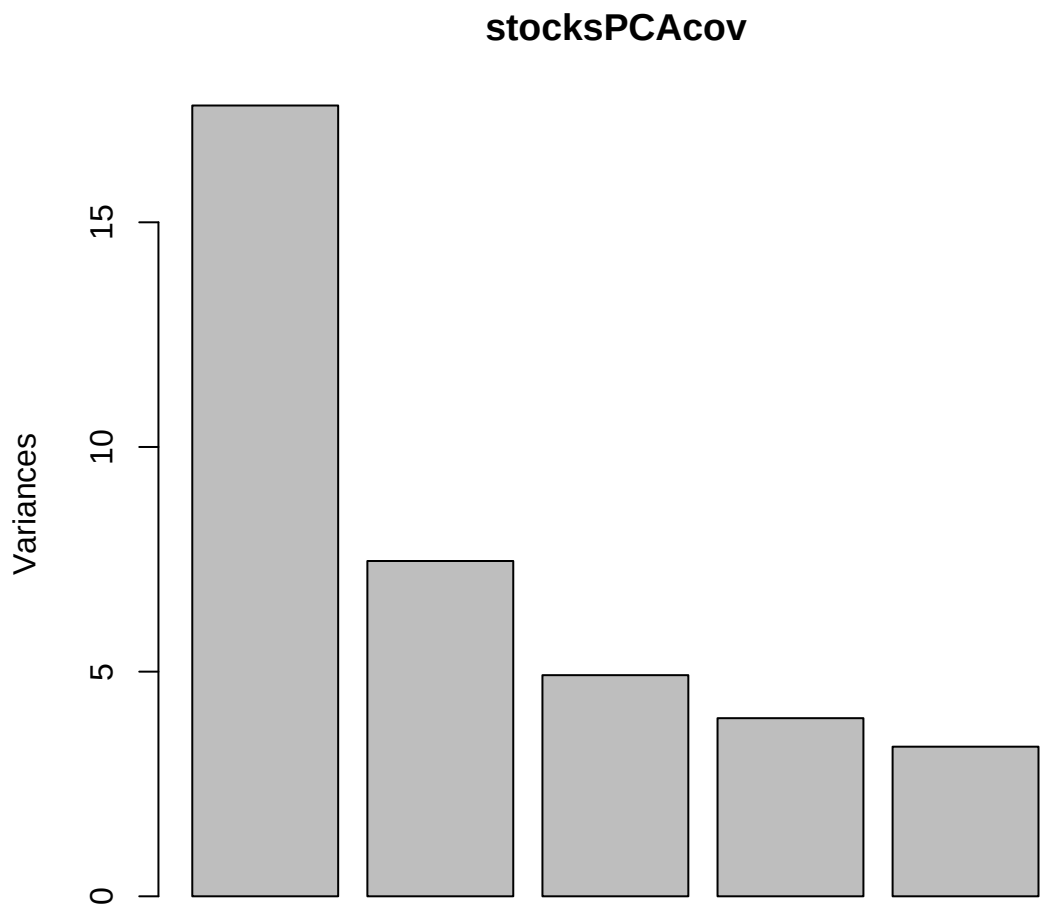
- R code and graphs for an updated set of stocks: DD, HON, and XOM, plus MSFT (Microsoft) and WMT (Walmart).

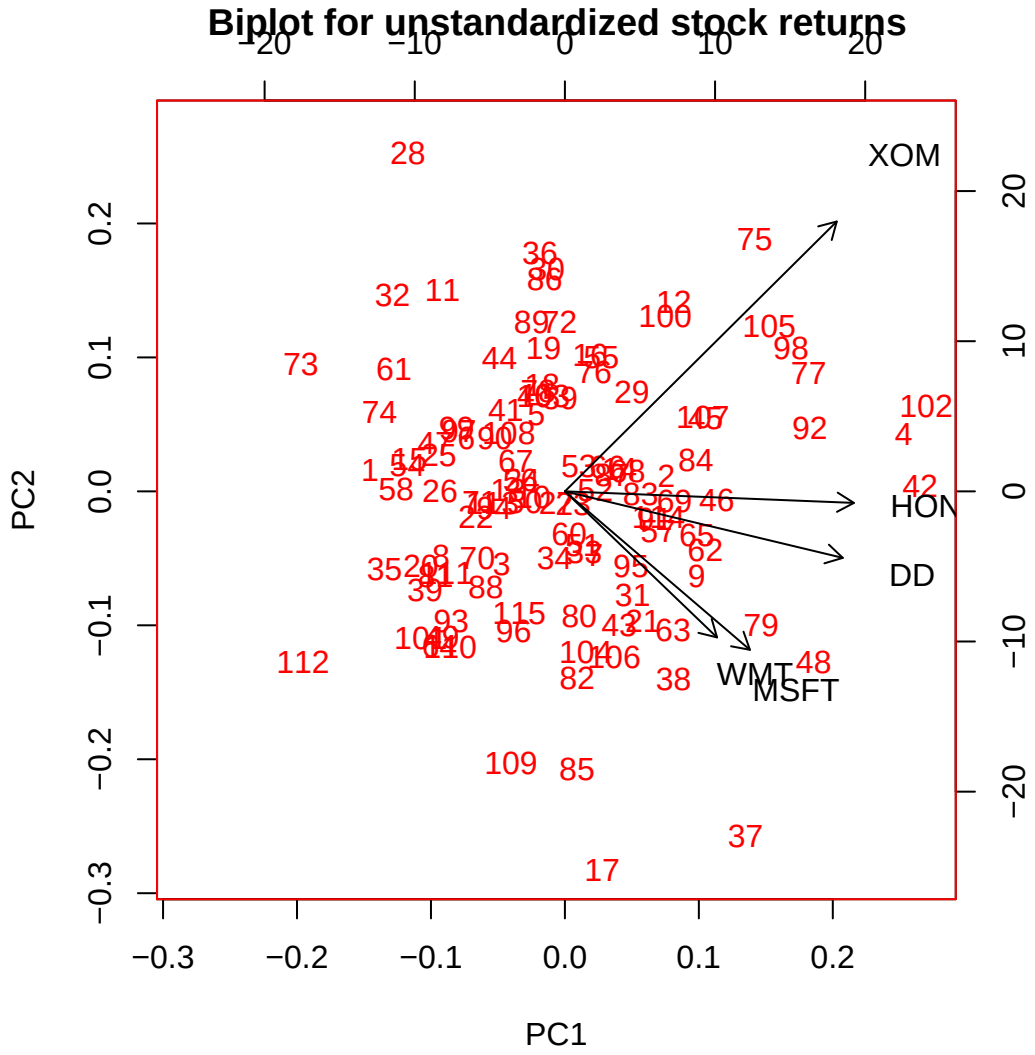
```
stocksPCAcor = prcomp(stocks(), scale. = TRUE);
print(stocksPCAcor);
plot(stocksPCAcor);
biplot(stocksPCAcor);

stocksPCAcov = prcomp(stocks());
print(stocksPCAcov);
plot(stocksPCAcov);
biplot(stocksPCAcov);
```


stocksPCAcor
Variances
0.0
0.5
1.0
1.5
2.0


Biplot for standardized stock returns
PC1
PC2
WMT
MSFT
DD
HON
XOM
−5
0
5
−0.2
−0.1
0.0
0.1
0.2


stocksPCAcov
Variances
0
5
10
15


Biplot for unstandardized stock returns
XOM
HON
DD
WMT
MSFT
PC1
PC2