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 correlation.

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

- Diagonal matrix: If

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

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 \( e_1 = p^{-1/2}(1, 1, \ldots, 1)' \);
- \( \lambda_k = 1 - \rho, \ k > 1 \);
- \( e_2, e_3, \ldots, e_p \) are an arbitrary basis for the rest of \( \mathbb{R}^p \).
- If \( \rho < 0 \) the order is reversed, but note that \( \rho \) must satisfy \( 1 + (p - 1)\rho \geq 0 \Rightarrow \rho \geq -1/(p - 1) \).
• Time series (1\textsuperscript{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

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

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

**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

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

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

– the coefficients \(V\);

– the values \(UD\).

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

\[ R = X^*/X^* \]

and the singular value decomposition of \(X^*\) gives the PCA of \(R\).
• 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);
```
stocksPCA cov

Variances
0 5 10 15