Principal Component Analysis

Max Turgeon

STAT 4690–Applied Multivariate Analysis

Population PCA i

- **PCA**: Principal Component Analysis
- Dimension reduction method:
  - Let $\mathbf{Y} = (Y_1, \ldots, Y_p)$ be a random vector with covariance matrix Σ. We are looking for a transformation $h : \mathbb{R}^p \to \mathbb{R}^k$, with $k \ll p$ such that $h(\mathbf{Y})$ retains "as much information as possible" about $\mathbf{Y}$.
- In PCA, we are looking for a **linear transformation** $h(y) = w^T y$ with **maximal variance** (where $\|w\| = 1$)

Population PCA ii

- More generally, we are looking for k linear transformations $w_1, \ldots, w_k$ such that $w_j^T \mathbf{Y}$ has maximal variance and is uncorrelated with $w_1^T \mathbf{Y}, \ldots, w_{j-1}^T \mathbf{Y}$.
- First, note that $\mathrm{Var}(w^T \mathbf{Y}) = w^T \Sigma w$. So our optimisation problem is

$$\max_w w^T \Sigma w, \quad \text{with } w^T w = 1.$$

- From the theory of Lagrange multipliers, we can look at the *unconstrained* problem

$$\max_{w,\lambda} w^T \Sigma w - \lambda(w^T w - 1).$$

Population PCA iii

- Write $\phi(w, \lambda)$ for the function we are trying to optimise. We have

$$\begin{aligned}\frac{\partial}{\partial w}\phi(w, \lambda) &= \frac{\partial}{\partial w} w^T \Sigma w - \lambda(w^T w - 1) \\ &= 2\Sigma w - 2\lambda w; \\ \frac{\partial}{\partial \lambda}\phi(w, \lambda) &= w^T w - 1.\end{aligned}$$

- From the first partial derivative, we conclude that

$$\Sigma w = \lambda w.$$

Population PCA iv

- From the second partial derivative, we conclude that $w \neq 0$; in other words, w is an eigenvector of Σ with eigenvalue λ.
- Moreover, at this stationary point of $\phi(w, \lambda)$, we have

$$\mathrm{Var}(w^T\mathbf{Y}) = w^T\Sigma w = w^T(\lambda w) = \lambda w^T w = \lambda.$$

- In other words, to maximise the variance $\mathrm{Var}(w^T\mathbf{Y})$, we need to choose λ to be the *largest* eigenvalue of Σ.
- By induction, and using the extra constraints $w_i^T w_j = 0$, we can show that all other linear transformations are given by eigenvectors of Σ.

Population PCA v

PCA Theorem
Let $\lambda_1 \geq \cdots \geq \lambda_p$ be the eigenvalues of Σ, with corresponding unit-norm eigenvectors $w_1, \ldots, w_p$. To reduce the dimension of $\mathbf{Y}$ from p to k such that every component of $W^T\mathbf{Y}$ is uncorrelated and each direction has maximal variance, we can take $W = \begin{pmatrix} w_1 & \cdots & w_k \end{pmatrix}$, whose j-th column is w_j.

Properties of PCA i

- Some vocabulary:
  - $\mathbf{Z}_i = w_i^T \mathbf{Y}$ is called the i-th **principal component** of $\mathbf{Y}$.
  - w_i is the i-th vector of **loadings**.
- Note that we can take $k = p$, in which case we do not reduce the dimension of $\mathbf{Y}$, but we *transform* it into a random vector with uncorrelated components.
- Let $\Sigma = P\Lambda P^T$ be the eigendecomposition of Σ. We have

$$\sum_{i=1}^p \mathrm{Var}(w_i^T \mathbf{Y}) = \sum_{i=1}^p \lambda_i = \mathrm{tr}(\Lambda) = \mathrm{tr}(\Sigma) = \sum_{i=1}^p \mathrm{Var}(Y_i).$$

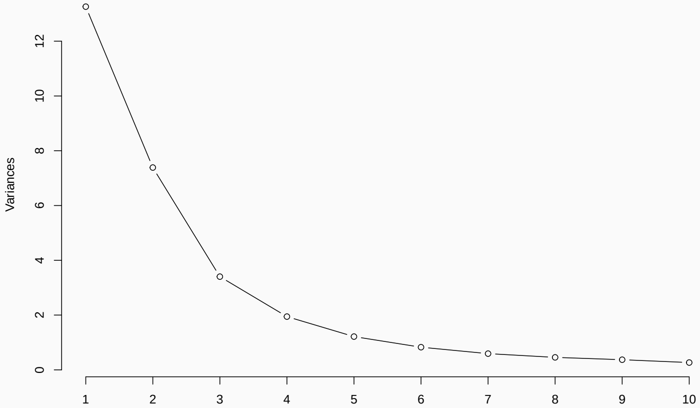
Properties of PCA ii

- Therefore, each linear transformation $w_i^T \mathbf{Y}$ contributes $\lambda_i / \sum_j \lambda_j$ as percentage of the overall variance.
- **Selecting** k: One common strategy is to select a threshold (e.g. $c = 0.9$) such that

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq c.$$

Scree plot

- A **scree plot** is a plot with the sequence $1, \dots, p$ on the x-axis, and the sequence $\lambda_1, \dots, \lambda_p$ on the y-axis.
- Another common strategy for selecting k is to choose the point where the curve starts to flatten out.
  - **Note**: This inflection point does not necessarily exist, and it may be hard to identify.


Variances
0
2
4
6
8
10
12
1
2
3
4
5
6
7
8
9
10

Correlation matrix

- When the observations are on the different scale, it is typically more appropriate to normalise the components of $\mathbf{Y}$ before doing PCA.
    - The variance depends on the units, and therefore without normalising, the component with the “smallest” units (e.g. centimeters vs. meters) will be driving most of the overall variance.
- In other words, instead of using Σ, we can use the (population) correlation matrix R.
- **Note**: The loadings and components we obtain from Σ are **not** equivalent to the ones obtained from R.

Sample PCA

- In general, we do not the population covariance matrix Σ.
- Therefore, in practice, we estimate the loadings w_i through the eigenvectors of the sample covariance matrix S_n.
- As with the population version of PCA, if the units are different, we should normalise the components or use the sample correlation matrix.

Example i

```
C <- chol(S <- matrix(c(1, 0.5, 0.1,
                        0.5, 1, 0.5,
                        0.1, 0.5, 1),
                      ncol = 3))

set.seed(17)
X <- matrix(rnorm(300), 100, 3)
Z <- X %*% C  ## ==>  cov(Z) ~=  C'C = S
pca <- prcomp(Z)
```

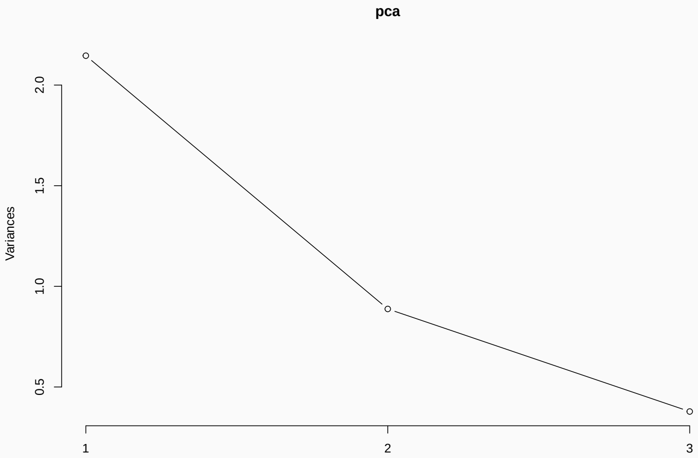
Example ii

```r
summary(pca)
```

```
## Importance of components:
##                          PC1    PC2    PC3
## Standard deviation     1.465 0.9422 0.6149
## Proportion of Variance 0.629 0.2602 0.1108
## Cumulative Proportion  0.629 0.8892 1.0000
```

```r
screeplot(pca, type = 'l')
```

Example iii



Example 2 i

```
pca <- prcomp(USArrests, scale = TRUE)
```

```
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.5749 0.9949 0.59713 0.41645
## Proportion of Variance 0.6201 0.2474 0.08914 0.04336
## Cumulative Proportion  0.6201 0.8675 0.95664 1.00000
```

Example 2 ii

```r
screeplot(pca, type = 'l')
```

Example 2 iii